

Empirical Null Estimation using Discrete Mixture Distributions and its Application to Protein Domain Data

Iris Ivy Gauran¹, Junyong Park¹, Johan Lim², DoHwan Park¹, John Zylstra¹, Thomas Peterson³, Maricel Kann³ and John Spouge⁴

¹Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, MD 21250, USA

²Department of Statistics, Seoul National University, Seoul, 08826, Republic of Korea

³Department of Biological Sciences, University of Maryland, Baltimore County, Baltimore, MD 21250, USA

⁴National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Abstract

In recent mutation studies, analyses based on protein domain positions are gaining popularity over gene-centric approaches since the latter have limitations in considering the functional context that the position of the mutation provides. This presents a large-scale simultaneous inference problem, with hundreds of hypothesis tests to consider at the same time. This paper aims to select significant mutation counts while controlling a given level of Type I error via False Discovery Rate (FDR) procedures. One main assumption is that there exists a cut-off value such that smaller counts than this value are generated from the null distribution. We present several data-dependent methods to determine the cut-off value. We also consider a two-stage procedure based on screening process so that the number of mutations exceeding a certain value should be considered as significant mutations. Simulated and protein domain data sets are used to illustrate this procedure in estimation of the empirical null using a mixture of discrete distributions.

Keywords: Local False Discovery Rate, Zero-Inflated Generalized Poisson, Protein Domain

1 Introduction

Interest towards multiple testing procedures has been growing rapidly in the advent of the so-called genomic age. With the breakthrough in large-scale methods to purify, identify and characterize DNA, RNA, proteins and other molecules, researchers are becoming increasingly reliant on statistical methods for determining the significance of biological findings ([40]). Gene-based analyses of cancer data are classic examples of studies which present thousands of genes for simultaneous hypothesis testing. However, [31] reported that gene-centric cancer studies are limited since the functional context that the position of the mutation provides is not considered. In lieu of this, [31] and [53] have shown that protein domain level analyses of cancer somatic variants could provide additional insights.

In particular, these studies can identify functionally relevant somatic mutations where traditional gene-centric methods fail by focusing on protein domain regions within genes, leveraging the modularity and polyfunctionality of genes. In protein domain-centric analyses of somatic mutations, somatic mutations from sequenced tumor samples are mapped from their genomic positions to positions within protein domains, enabling the comparison of distant genomic regions that share similar structure and amino acid composition ([36]; [37]; [38]). In the analysis of sequenced tumor samples, it is assumed that the mutational distribution will consist of many “passenger” mutations, which are non-functional randomly distributed background mutations, in addition to rare functional “driver” mutations that reoccur at specific sites within the domain and contribute to the initiation or progression of cancer ([35]; [47]; [46]). The problem that is addressed here is in a single domain, how to identify the highly mutated positions compared to the background where the number of positions in a domain can be as large as several tens or hundreds.

Motivated by the aforementioned domain-level analyses, we propose a methodology for identifying significant mutation counts while controlling the rate of false rejections. [9] reported that much of the statistics microarray literature is focused on controlling the probability of a Type I error, a “false discovery”. A traditional approach is to control the family-wise error rate (FWER), the probability of making at least one false discovery. However, with the collection of simultaneous hypothesis tests in the hundreds or thousands, trying to limit the probability of even a single false discovery leads to lack of power. Alternatively, in a seminal paper, [4] introduced a multiple-hypothesis testing error measure called False Discovery Rate (FDR). This quantity is the expected proportion of false positive findings among all the rejected hypotheses. Among the FDR-controlling test methods, [14] developed an empirical Bayes approach where they established a close connection between the estimated posterior probabilities and a local version of the FDR.

A key step in controlling the local false discoveries is to estimate the null distribution of the test statistics. [10] stated that the test statistics in large-scale testing may not accurately follow the theoretical null distribution. Instead, the density of the null distribution is estimated from the large number of genes. In these microarray experiments, [11] employed a normal mixture model and proposed maximum likelihood and mode matching to estimate the empirical null distribution. Using the same normal mixture model, [20] proposed a method to estimate the empirical null based on characteristic functions. In addition, [34] proposed a local FDR estimation procedure based on modeling the null distribution with a mixture of normal distributions. However, these existing methods are based on the assumption that the null is a mixture of continuous distributions. In the case of domain-level analyses, the data is characterized as mutation counts among N positions in the domain. This indicates that the available methods in the estimation of the empirical null should be extended to a mixture of discrete distributions.

The rest of the paper is organized as follows. In Section 2, we discuss the problem in detail and review two existing multiple testing procedures, namely Efron’s Local FDR procedure and Storey’s procedure. In Section 3, we introduce the estimation procedure for f_0 , f and π_0 , where the null distribution is assumed to be a zero-inflated model. Also, a novel two-stage multiple testing procedure is presented in this section. In Section 4, the performance of the new procedure is studied via simulations and the results for real data sets are presented. Some concluding remarks will be presented in Section 5.

2 Multiple Testing Procedures controlling FDR

In this section, we briefly discuss the motivating example and review the existing procedures for analysis. The collection of the original dataset is $\mathbf{a} = (a_1, a_2, \dots, a_N)'$, where a_i is the number of mutations in the i th position of the specific domain with N positions. We define $\mathcal{A} = \{j : j \geq 0, n_j > 0\}$ as the set of the unique values of \mathbf{a} , $K = \max(\mathbf{a})$, and L is the cardinality of \mathcal{A} where $L \leq K + 1$. Some relevant features of \mathbf{a} follow. A large proportion of positions do not have any mutation, $a_i = 0$. Also, L is relatively small compared to N , which means that the number of mutations in many positions are tied. Since our goal is to identify the positions with extra disease mutation counts, it is only reasonable to have the same conclusion for positions wherein the number of mutations are tied. Therefore, we transform the data into the observed “histograph” of positions over “mutation counts”. We define $n_j = |\{i : a_i = j\}|$, as the number of positions with j mutations, $j \in \mathcal{A}$, and $\sum_{j=0}^K n_j = N$. The ordered data \mathbf{x}_N can be represented as a partition of the unique values of \mathbf{a} , that is,

$$\mathbf{x}'_N = (\mathbf{x}'_0, \mathbf{x}'_1, \dots, \mathbf{x}'_K) = (\underbrace{0, 0, \dots, 0}_{\mathbf{x}'_0}, \underbrace{1, 1, \dots, 1}_{\mathbf{x}'_1}, \dots, \underbrace{K, K, \dots, K}_{\mathbf{x}'_K})$$

where \mathbf{x}_j is the column vector containing n_j of j s. Since the information contained in \mathbf{x}_j is analogous to knowing n_j , for any $j \in \mathcal{A}$, then another, equivalent format of the data set is $\mathbf{y}_N = (n_0, n_1, \dots, n_K)'$.

For any single domain of interest, a total of L mutation counts can be decomposed into two groups, \mathcal{A}_0 and \mathcal{A}_1 , where \mathcal{A}_0 is the collection of small number of mutation counts which is considered to be non-significant and \mathcal{A}_1 is the set of large number of mutation counts which consists of significantly mutated positions. Let the prior probabilities of the two groups be π_0 or $\pi_1 = 1 - \pi_0$, and assume corresponding densities, f_0 or f_1 . Define f_0 to be the null distribution and f_1 to be the alternative distribution. Therefore, we consider the problem of testing L null hypotheses simultaneously,

$$H_0 : H_{0j} \text{ is true for } j \in \mathcal{A}$$

on the basis of a data set \mathbf{a} , where H_{0j} is stated as the number of mutations j is generated from f_0 for all $j \in \mathcal{A}$ with $|\mathcal{A}| = L$. For a given position, the number of mutations follow one of the two distributions f_0 or f_1 , so the probability density function of the mixture distribution can be represented as

$$f(a) = \pi_0 f_0(a) + (1 - \pi_0) f_1(a) \tag{1}$$

and our goal is to identify the positions which have significantly different patterns from the null.

For continuous data, [11] introduced the idea of “zero assumption” where observations around the central peak of the distribution consists mainly of null cases. Using this assumption, f_0 is estimated using Gaussian quadrature which is based on derivative at the mode. However, such a procedure is not applicable to discrete data. In our problem on discrete data, we introduce the following assumption on the null distribution which plays a key role throughout this paper.

Assumption on f_0 :

$$f_1(a) = 0 \text{ for } a \leq C \text{ for some } C \in \mathbb{Z}^+. \quad (2)$$

From the assumption, $a_i \leq C$ are guaranteed to be from f_0 and $a_i > C$ are generated from the mixture of f_0 and f_1 . For notational convenience, we relabel the data as $\mathbf{x}_n = (\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_C)$ for the null sample and $\mathbf{x}_{N-n} = (\mathbf{x}_{C+1}, \mathbf{x}_{C+2}, \dots, \mathbf{x}_K)$ for the mixture of null and non-null samples. The sampling distribution for the null sample is f_0 itself while f in (1) is the sampling distribution of the non-null sample. We will discuss more details about how to choose the value of C in the next section.

Following the pioneering work of [4], we employ the sequential p-value method to determine r that tells us to reject $p_{(1)}, p_{(2)}, \dots, p_{(r)}$, where $p_{(1)}, p_{(2)}, \dots, p_{(K)}$ are the ordered observed p-values. [45] improved the Benjamini-Hochberg procedure with the inclusion of the estimator of the null proportion, $\hat{\pi}_0$, which indicates that we reject $p_{(1)}, p_{(2)}, \dots, p_{(l)}$ such that

$$l = \max \left\{ i : p_{(i)} \leq \frac{\alpha \sum_{j \geq i} n_j}{N \hat{\pi}_0} \right\} \quad (3)$$

The BH procedure and Storey's procedure are equivalent, that is $r = l$, if we take $\hat{\pi}_0 = 1$. The details about the estimation of π_0 is provided in the next section. Moreover, following [13], we define the local FDR at any mutation count, say t , as

$$fdr(t) = \frac{\pi_0 f_0(t)}{f(t)} \quad (4)$$

which indicates that $fdr(t)$ is the posterior probability of a true null hypothesis at t . The interpretation of the local FDR value is analogous to the frequentist's p-value wherein local FDR values less than a specified level of significance provide stronger evidence against the null hypothesis.

3 Methodology

3.1 Model Specification

Depending on the application, we assume that the mutation counts follow a zero-inflated model in order to account for the true zeros in the count model and the excess zeros. The class of models considered is the Generalized Poisson (GP) distribution introduced by [6], with an additional zero-inflation parameter.

Let T be a nonnegative integer-valued random variable where relative to Poisson model, it is overdispersed with variance to mean ratio exceeding 1. If $T \sim GP(\lambda, \theta)$, then the probability mass function can be written as

$$P(T = t) = g(t) = \frac{\lambda(\lambda + \theta t)^{t-1}}{t!} e^{-\lambda - \theta t} \quad (5)$$

where $0 \leq \theta < 1$ and $\lambda > 0$.

If zero is observed with a significantly higher frequency, we can include a zero-inflation parameter to characterize the distribution. Then $X \sim ZIGP(\eta, \lambda, \theta)$ and the probability that $X = j$, denoted by $f_0(j)$, is

$$f_0(j) = \begin{cases} \eta + (1 - \eta)e^{-\lambda} & j = 0 \\ (1 - \eta)g(j) & j = 1, 2, \dots \end{cases}$$

where j is a nonnegative integer, $0 \leq \eta < 1$, $0 \leq \theta < 1$ and $\lambda > 0$. Recently, ZIGP models have been found useful for the analysis of heavy-tailed count data with a large proportion of zeros ([17]; [15]; [16]). The ZIGP model reduces to Zero-Inflated Poisson (ZIP) distribution when $\theta = 0$, Generalized Poisson distribution (GP) when $\eta = 0$ and Poisson distribution when $\eta = 0$ and $\theta = 0$.

The ZIP model, first introduced by [25], is applied when the count data possess the equality of mean and variance property while taking into consideration the structural zeros and zeros which exist by chance. Meanwhile, the Zero-Inflated Negative Binomial (ZINB) model is widely used for handling data with population heterogeneity which may be caused by the occurrence of excess zeros and the overdispersion due to unobserved heterogeneity ([39]). Several studies show that ZINB model provides a better fit to the overdispersed count data when ZIP is inadequate ([52]; [50]; [23]). However, [21] showed that the ZIGP distribution provides a better fit than ZINB when there is a large fraction of zeros and the data is heavily right-skewed. They compared the probabilistic properties of the zero-inflated variations of NB and GP distributions, such as probability mass and skewness, while keeping the first two moments fixed. Using this result, it is worthwhile to consider ZIGP rather than ZINB given that the mutation count data exhibited both features.

3.2 Estimation of f_0 , f and π_0

From (4), the local FDR formulation consists of unknown quantities f_0 , f , and π_0 which must be estimated accordingly. We follow the idea of “zero assumption” in [11] which modeled f_0 to normal null and [34] which modeled f_0 as a mixture of normals. In the proposed method, we apply f_0 to the context of ZIP and ZIGP models which indicate that a small mutation count suggests a few random background mutations, whereas a large mutation count suggests a mixture of a few background and a lot of functional disease mutations. However, since f_0 is unknown in practice, four count models will be compared in order to come up with estimates for the parameters of the null distribution. These models belong to the class of ZIGP distribution, namely, (1) ZIGP (2) ZIP (3) Generalized Poisson and (4) Poisson. If the true f_0 is ZIGP and the model used to estimate f_0 is ZIGP then we expect superior results compared to the other three distributions. Moreover, if the true null distribution is ZIP, then we expect better results for ZIP and ZIGP distribution compared to GP and Poisson distribution. This suggests that since ZIGP can characterize overdispersion, even if there is none such as the case of ZIP, it should still be able to capture the behavior of f_0 accurately.

To estimate the parameters of f_0 for any of these four count models, the EM Algorithm proposed by [28] will be utilized. For truncated data sets described in (2), fitting the model using EM algorithm is not straightforward as when all data points are available. In general, the M-step of this algorithm does not have a closed form unless the complete data vector is extended to include indicator variables denoting the membership of data points with respect to the components of the mixture.

If the null distribution is assumed to be ZIGP, then the log likelihood $\ell(\eta, \lambda, \theta \mid \mathbf{x}_N)$ of the entire data vector is

$$\sum_{j=0}^C n_j \log f_0(j; \Theta) + \sum_{j=C+1}^K n_j \log f(j; \cdot) \quad (6)$$

Suppose the sample space of X , denoted by \mathcal{X} , is partitioned into $K+1$ mutually exclusive subsets $\mathcal{X}_j = \{j\}$, $j \in \mathcal{A}$, where independent observations are made on X . After choosing a suitable value for C , the null sample $\mathbf{x}_n = (\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_C)$ and the corresponding vector of mutation counts $\mathbf{y}_n = (n_0, n_1, \dots, n_C)'$ are available for the estimation of the parameters of f_0 . However, the problem that arises is that the number of observations n_j falling in \mathcal{X}_j , $j > C$ are not available for the subsequent estimation of the parameters of f_0 .

For the n observations in \mathbf{x}_n , it is assumed that $\mathbf{y}_n = (n_0, n_1, \dots, n_C)'$ has a Multinomial distribution consisting of n draws on $C+1$ categories with probabilities p_j

$$p_j = \frac{f_0(j; \Theta)}{\sum_{j=0}^C f_0(j; \Theta)} \quad (7)$$

where $\Theta = (\eta, \lambda, \theta)$, $\sum_{j=0}^C p_j = 1$ and $\sum_{j=0}^C n_j = n$. This gives the likelihood function

$$L_0(\Theta; \mathbf{y}_n) = \frac{n!}{n_0! n_1! \dots n_C!} \prod_{j=0}^C p_j^{n_j} \quad (8)$$

From (8), we can solve the likelihood equation $\partial L_0(\Theta; \mathbf{y}_n) / \partial \Theta = \mathbf{0}$ within the EM framework following the work of [8]. The EM machinery is invoked by defining $\mathbf{w}_N = (\mathbf{y}'_n, \mathbf{y}'_{N-n})'$ as the complete-data vector where $\mathbf{y}_{N-n} = (n_{C+1}, n_{C+2}, \dots, n_K)'$. Then, instead of looking at the log likelihood for \mathbf{y}_n , we consider the log likelihood function of the complete data, $\ell(\eta, \lambda, \theta | \mathbf{w}_N)$. In order to find the estimates, it is important to note that each entry of \mathbf{y}_{N-n} is a realization of a hidden random variable. However, since these realizations do not exist in reality, we have to consider each entry of \mathbf{y}_{N-n} as a random variable itself.

Furthermore, [28] proposed an extension of the complete-data vector \mathbf{w}_N for mixture densities to include the zero-one indicator variables

$$\mathbf{z}_{jk} = (z_{0jk}, z_{1jk})' \quad j = 0, 1, \dots, K; k = 1, 2, \dots, n_j$$

where $z_{0jk} + z_{1jk} = 1$ and given the number of mutations j , \mathbf{z}_{jk} are conditionally independent. Conditional on the value of j , the probability of membership to a component can be computed using Bayes' Theorem as

$$\tau_{0j}(\Theta) = P(z_{0jk} = 1 | j) = \frac{\eta I_{\{0\}}(j)}{f_0(j)}$$

and $\tau_{1j}(\Theta) = P(z_{1jk} = 1 | j) = 1 - P(z_{0jk} = 1 | j)$. The indicator function $I_S(j)$ is equal to 1 if $j \in S$ and 0 otherwise.

Using these indicator variables in the complete-data specification, the log likelihood becomes

$$\sum_{j=0}^K \sum_{k=1}^{n_j} z_{0jk} \log \eta I_{\{0\}}(j) + \sum_{j=0}^K \sum_{k=1}^{n_j} z_{1jk} \log [(1 - \eta)g(j)] \quad (9)$$

The details of the EM Algorithm are provided in the Appendix. Moreover, it is straightforward to estimate $f(j)$ by using relative frequency given by

$$\hat{f}(j) = \frac{n_j}{n_0 + n_1 + \dots + n_K} \quad (10)$$

Using the assumption on f_0 , for $j \leq C$, $f(j)$ from (1) reduces to $\pi_0 f_0(j)$. Hence,

$$\sum_{j=0}^C \pi_0 f_0(j) = \sum_{j=0}^C f(j)$$

To estimate π_0 , we need to calculate

$$\hat{\pi}_0 = \frac{\sum_{j=0}^C \hat{f}(j)}{\sum_{j=0}^C \hat{f}_0(j)} \quad (11)$$

using (10) and the estimate of f_0 after plugging in $\hat{\Theta}$ resulting from EM algorithm. Finally, the estimate of π_0 is $\min(1, \hat{\pi}_0)$.

3.3 Choice of the Cut-off C

In our model, we assume that we can identify a cut-off C , wherein bins with number of mutations greater than C contain more mutations than what would be expected in the null model. The choice of the cut-off C is of paramount importance since the estimation of f_0 and π_0 depend on C . It is more realistic to assume that C is unknown, so such a predetermined C may affect the result of local FDR procedure seriously.

In particular, if C is predetermined and is chosen to be larger than the true value, the null distribution is estimated based on observations from alternative hypothesis as well as null hypothesis, so the estimated null distribution is contaminated by the alternative distribution. This will cause insensitivity of local FDR procedure in detecting the alternative hypothesis. On the other hand, if C is chosen to be smaller, then the null distribution is estimated only based on small values, so the estimation of the null distribution especially at the tail part is less reliable. Empirically, the FDR procedure yields liberal results in that there are too many rejections resulting in failure in controlling a given level of FDR.

For the normal distribution as a null distribution, [12] proposed the maximum likelihood estimation from likelihood based on observations in a given predetermined interval around zero. [34] considered a mixture of normal distributions for the null distribution and proposed two approaches to select intervals around the mode to estimate the parameters in the mixture model using the EM algorithm. One of the proposed methods is based on the idea of goodness of fit to the parametric model of the null distribution. As the interval increases in length, it finally includes more and more alternative values resulting in deviation from the null distribution.

The estimation of the cut-off C has been also formulated in the context of change-point analysis. [44] offer an objective-change point method that can replace the subjective approaches performed by eye-balling the data. The proposed method resembles the change-point regression and robust regression but it is tailored to estimate the change point from a transient to an asymptotic regime. Given a tuning parameter c and a criterion function ρ , depending on β , the estimator for the change point k^* is defined as

$$k^* = \arg \min_{k=0,1,\dots,n} \left(\min_{\beta} \sum_{i=k+1}^n (\rho(e_i) - c) \right) \quad (12)$$

where $\rho(e_i)$ is the estimated least-squares normalized residual. In (12), there is a tuning parameter c which should be given ahead. The value of c plays the role of penalty for adding terms $\rho(e_i)$ in (12), so the predetermined value of c affects k^* arbitrarily. We see that our proposed estimation of C is related to the form (12).

We introduce our proposed estimation procedure of C . Let us define the index sets

$$\mathcal{A} = \{j : j \geq 0, n_j > 0\}, \quad \mathcal{A}(C) = \{j : 0 \leq j \leq C, n_j > 0, f_1(j) = 0\}. \quad (13)$$

Note that $\mathcal{A}(C_1) \subset \mathcal{A}(C_2)$ for $C_1 < C_2$ and $f(j) = \pi_0 f_0(j)$ when $j \in \mathcal{A}(C)$. We adopt the idea of sequential testing to detect the change point in which the observations are generated from the mixture distribution f . More specifically, suppose we observed $(0, n_0), (1, n_1), \dots, (K, n_K)$ sequentially from $f_0(0), f_0(1), \dots, f_0(C), f(C+1), \dots, f(K)$ where distribution is changed from f_0 to f at $C+1$.

Our goal is to detect the change point C based on assuming that we observe $0, 1, 2, \dots, K$ sequentially. For a given ν , we define $S_\nu(\Theta)$ as

$$S_\nu(\Theta, f) = \sum_{j \leq \nu} n_j \log f_0(j) + \sum_{j \geq \nu+1} n_j \log f(j) = \sum_{j \leq \nu} n_j \log \frac{f_0(j)}{f(j)} + \sum_{j \leq K} n_j \log f(j). \quad (14)$$

Maximizing $S_\nu(\Theta, f)$ is equivalent to the CUSUM(cumulative sum) $\sum_{j \leq \nu} n_j \log \frac{f_0(j)}{f(j)}$. Since the parameters Θ is estimated from EM algorithm and $\hat{f}(j) = \frac{n_j}{M}$, our procedure is

$$\hat{C} = \operatorname{argmax}_{\nu=1,2,\dots,K} S_\nu(\hat{\Theta}_\nu) \quad (15)$$

where $\hat{\Theta}_\nu$ is the estimator from the EM algorithm in the previous section with the value of C set to ν . One may consider the full likelihood of all observations and find out some connection between S_ν and the full likelihood presented as follows.

The likelihood function of $(0, n_0), \dots, (K, n_K)$ for a given $\mathcal{A}(\nu)$ is

$$\text{likelihood} \equiv L(\Theta^*, f) = \prod_{j \leq K} f(j)^{n_j} = \prod_{j \leq \nu} (\pi_0 f_0(j))^{n_j} \prod_{j \geq \nu+1} f(j)^{n_j} \quad (16)$$

where $\pi_0 f_0$ depends on $\Theta^* = (\pi_0, \eta, \theta, \lambda) = \{\pi_0\} \cup \Theta$. The log likelihood is also

$$\log L(\Theta^*, f) = \ell_\nu(\Theta^*, f) \equiv \sum_{j \leq \nu} n_j \log(\pi_0 f_0(j)) + \sum_{j \geq \nu+1} n_j \log f(j) \quad (17)$$

since $f(j) = \pi_0 f_0(j)$ for $j \in \mathcal{A}(\nu)$. This leads to

$$\ell_\nu(\Theta^*, f) \equiv \sum_{j \leq \nu} n_j \log \frac{\pi_0 f_0(j)}{f(j)} + \sum_{j \leq K} n_j \log f(j) \quad (18)$$

which is equivalent to

$$S_\nu(\Theta^*, f) = \ell_\nu(\Theta^*, f) - N_\nu \log \pi_0 \quad (19)$$

$$= \sum_{j \leq \nu} n_j (\text{lr}_j(\Theta^*, f) - \log \pi_0) + l_0 \quad (20)$$

where $N_\nu = \sum_{j \leq \nu} n_j$ and $\text{lr}_j(\Theta^*, f) = \frac{\pi_0 f_0(j)}{f(j)}$. It can be also seen that the penalized likelihood has the form of (12)

$$\sum_{j \in \mathcal{A}(C)} n_j (-\text{lr}_j(\Theta^*, f) - c) \quad (21)$$

where $c = -\log \pi_0$. We estimate C via

$$\hat{C}_1 = \text{argmin}_{\nu=1,2,\dots,K} \left(-S_\nu(\hat{\Theta}_\nu^*, \hat{f}) \right) = \text{argmin}_{\nu=1,2,\dots,K} \left(-\ell_\nu(\hat{\Theta}_\nu^*, \hat{f}) + N_\nu \log \hat{\pi}_{0,\nu} \right) \quad (22)$$

$$= \text{argmin}_{\nu=1,2,\dots,K} \sum_{j \in \mathcal{A}(C)} n_j \left(\rho_j(\hat{\Theta}_\nu^*, \hat{f}) - \hat{c}_\nu \right) \quad (23)$$

where $\hat{\Theta}_\nu^* = (\hat{\pi}_{0,\nu}, \hat{\eta}_\nu, \hat{\theta}_\lambda, \hat{\lambda}_\nu)$ is obtained from the EM algorithm discussed in the previous section, $\hat{f}(j) = n_j/N$, $\rho_j(\hat{\Theta}_\nu^*, \hat{f}) = -\text{lr}_j(\hat{\Theta}_\nu^*, \hat{f})$ and $\hat{c}_\nu = -\log \hat{\pi}_{0,\nu}$.

In (12), c is a predetermined value, however we don't need to predetermine any parameter in (23). The proposed criterion (23) is related to the penalized model selection such as AIC and BIC. When we use the information that $n = \sum_{j \leq C} n_j$ observed values are generated from f_0 , $\sum_{j \leq \nu} n_j \text{lr}_j(\Theta^*, f)$ is increasing in ν , there is a compromise term $c = -\log \pi_0$ for each observation to compensate adding additional terms. There is a total of N_ν positions, so when we use the assumption $\nu = C$, we consider $N_\nu \log \pi_0$ penalty to the log likelihood function ℓ_ν . Most of well known model selection criteria have similar forms where the penalty terms are related to penalize the complexity of models. In our context, the term $-\log \pi_0$ gives penalty to using the information that j for $j \leq \nu$ are generated from f_0 . For a small value of π_0 , the corresponding penalty ($-\log \pi_0$) is large since a large penalty should be given to a low chance of f_0 . On the other hand, if π_0 is close to 1, there becomes small risk from assuming observations are from the null hypothesis.

For the second method, we consider the extension of the methodology proposed by [12] which explicitly uses the zero assumption. This stipulates that the non-null density f_1 is supported outside some set $\{0, 1, \dots, C\}$. Let n be the number of mutations which is at most C and define the likelihood function for \mathbf{x}_n as

$$L(\hat{\Theta}_\nu^\star | \mathbf{x}_n) = \xi^n (1 - \xi)^{N-n} \prod_{j \leq \nu} (f_0(j))^{n_j}$$

where $\xi = \hat{\pi}_0 \sum_{j=0}^C \hat{f}_0(j)$. The cut-off can be computed as

$$\hat{C}_2 = \operatorname{argmin}_{\nu=1,2,\dots,K} \left(\log L(\hat{\Theta}_\nu^\star | \mathbf{x}_n) \right) \quad (24)$$

3.4 Modification of local FDR by truncation

In practice, if a given domain position has a large number of mutations, then these mutations are expected to be significant. In many cases, there are relatively few positions in a protein domain where large values of mutations can be observed. This indicates that for large values of j , estimation of f based on relative frequency is not accurate due to the sparse data in the tail part. Consequently, the estimated local FDR is not reliable since it depends on the estimator of f .

Rather than testing significance based on inaccurate local FDRs from large mutation counts, we consider a screening process so that the number of mutations exceeding a certain value should be considered as significant mutations. Such a critical value will be decided depending on the estimated null distribution. When we have observations a_i for $1 \leq i \leq N$ generated from the null distribution, we are interested in figuring out D_N such that

$$P \left(\max_{1 \leq i \leq N} a_i < D_N \right) \rightarrow 1 \quad (25)$$

as $N \rightarrow \infty$. Once a sequence D_N is identified, $a_i (\geq D_N)$ is hardly observed under the null hypothesis, so the corresponding null hypothesis is rejected directly rather than making decision based on local FDR procedure. There are many choices of D_N , but a smaller sequence of D_N satisfying (25) is of our interest since any sequence B_N satisfying $B_N > D_N$ also satisfies the property.

When a_i is observed from Generalized Poisson distribution, [24] showed that the tail probabilities satisfy the following inequality:

$$P(a_i \geq D_N) < \left[1 - e^{1-\theta} \left(\theta + \frac{\lambda}{D_N + 1} \right) \right]^{-1} \frac{\lambda(\lambda + \theta D_N)^{D_N-1}}{(D_N)^{D_N+1/2}} e^{-\lambda - (\theta-1)D_N} \quad (26)$$

where $D_N \geq \frac{\lambda}{e^{\theta-1} - \theta}$, $\theta \in (0, 1)$, $\lambda > 0$.

Using (26), we can compute for

$$\begin{aligned} P\left(\max_{1 \leq i \leq N} a_i \geq D_N\right) &= 1 - [1 - P(a_i \geq D_N)]^N \\ &\leq 1 - \left[1 - (\delta_{D_N})^{-1} \frac{\lambda(\lambda + \theta D_N)^{D_N-1}}{(D_N)^{D_N+1/2}} e^{-\lambda - (\theta-1)D_N}\right]^N \end{aligned}$$

where $\delta_{D_N} = 1 - e^{1-\theta} \left(\theta + \frac{\lambda}{D_N+1}\right)$. For (25) to hold,

$$\log N - \log \delta_{D_N} + \log \left(\frac{\lambda(\lambda + \theta D_N)^{D_N-1}}{(D_N)^{D_N+1/2}} e^{-\lambda - (\theta-1)D_N} \right) \rightarrow -\infty \quad (27)$$

and (27) can be simplified in terms of N and D_N as

$$\mathcal{G}_N \equiv \log N - 0.5 \log D_N - \log(D_N + 1) + D_N \log \left(\theta + \frac{\lambda}{D_N} \right) - (\theta - 1)D_N$$

leading to

$$\mathcal{G}_N \asymp \log N + (\log \theta - \theta + 1)D_N. \quad (28)$$

To assure that $\mathcal{G}_N \rightarrow -\infty$, we can take $D_N = \zeta \log N$ for some constant ζ satisfying

$$\zeta > \frac{1}{\theta - 1 - \log \theta}$$

Since $\log \theta \leq \theta - 1$, then $\zeta > 0$, $\theta \in (0, 1)$ as desired. Hence, we take

$$D_N = \left\lceil \max \left(\frac{\lambda}{e^{\theta-1} - \theta}, \frac{\log N}{\theta - 1 - \log \theta} \right) \right\rceil \quad (29)$$

where $\lceil x \rceil$ is the smallest integer greater than or equal to x ($x > 0$). Meanwhile, if a_i is observed from Poisson distribution, [29] derived the bounds for the tail probabilities using the Chernoff bound argument:

$$P(a_i \geq D_N) < \frac{e^{-\lambda}(e\lambda)^{D_N}}{(D_N)^{D_N}} \quad (30)$$

where $0 < \lambda < D_N$. Using the inequality in (30),

$$P\left(\max_{1 \leq i \leq N} a_i \geq D_N\right) \leq 1 - \left(1 - \frac{e^{-\lambda}(e\lambda)^{D_N}}{(D_N)^{D_N}}\right)^N$$

and in order to satisfy the condition in (25), $\mathcal{P}_N \rightarrow -\infty$ where

$$\mathcal{P}_N \asymp \log N - D_N \log D_N$$

Therefore, we take

$$D_N = \lceil \max(\lambda, \log N) \rceil \quad (31)$$

When a_i is observed from ZIGP, D_N can be calculated exactly as shown in (29) since the derivation will eventually yield the leading terms in (28) which does not involve η . Similarly, if a_i is observed from ZIP, D_N can be computed using (31).

3.5 Two Stage Procedure

The proposed method can be summarized into two stages:

1. Using the likelihood method specified in Section 3.3, identify the cut-off point C .
2. Suppose $\hat{\Theta} = (\hat{\eta}, \hat{\lambda}, \hat{\theta})$ are the parameter estimates at the chosen C . Using $\hat{\Theta}$, compute D_N based on the specified formula in Section 3.4. By construction, we expect the value of D_N to fall within the interval $C < D_N \leq K$. However, it is probable to observe values of D_N outside this interval. Under these scenarios, we consider the following:
 - (a) If the calculated value of D_N exceeds K , we take $D_N = K$. This implies that there is no screening process performed.
 - (b) If the calculated value of D_N is below C , we take $D_N = C + 1$. This indicates that all values above the chosen C are automatically declared as significant mutations.

To incorporate these conditions on the formulation of D_N , we can modify (29) as

$$D_N = \min \left(\left\lceil \max \left(\frac{\lambda}{e^{\theta-1} - \theta}, \frac{\log N}{\theta - 1 - \log \theta}, C + 1 \right) \right\rceil, K \right) \quad (32)$$

and (31) as

$$D_N = \min (\lceil \max (\lambda, \log N, C + 1) \rceil, K) \quad (33)$$

For a given null distribution, we can calculate D_N using (32) or (33) correspondingly. After determining the value of D_N , all values of $j \geq D_N$ are considered significant mutations.

Using this two-stage procedure, we can identify the mutation counts which are falsely rejected. In the simulated data set, we can specify the value of true C . As discussed previously, all mutation counts below C are assumed to follow the null distribution f_0 . Hence, any rejection for mutation counts $j \leq C$ are considered to be erroneous.

4 Numerical Studies

4.1 Simulation Studies

To gain insights regarding the robustness of the proposed procedures in the presence of model misspecification, we perform some simulation studies. The comparison is based on four simulation boundaries: (1) method used in the choice of the cut-off C ; (2) model for the estimation of f_0 ; (3) null distribution; and (4) non-null distribution used in data generation. There are two methods considered for the choice of cut-off C as discussed in the previous section. The null distributions considered are Zero-Inflated Poisson (ZIP) and Zero-Inflated Generalized Poisson (ZIGP) distribution. Both distributions account for the excessive number of zeros which is a characteristic of the mutation count data. For the non-null distribution, Geometric($p = 0.08$) and Binomial($n = 250, p = 0.20$) distribution are utilized. These were chosen because it can characterize the pattern of the mutation count observed in the real data set.

The assessment of the performance is also based on the model used in the estimation of f_0 since it affects calculation of the local FDR. The four models compared are ZIGP, ZIP, Generalized Poisson and Poisson distribution. This allows for the comparison of the number of falsely rejected hypotheses when the model for f_0 is specified correctly and when there is departure from the true model of f_0 .

A total of L hypotheses tests were performed for independent random variables n_j over 1000 replications. For each replication, the proportion of n_j from the null distribution is set to be π_0 and the total number of positions N is specified to be 1000. To calculate the False Discovery Rate, \widehat{FDR} , for the k th generated data, $k = 1, 2, \dots, 1000$, we compute the false discovery proportion (FDP) which is defined by

$$FDP_k = \frac{V_k}{R_k} I(R_k > 0)$$

where V_k and R_k are the number of falsely rejected hypotheses (false discoveries) and the total number of rejected hypotheses in the k th generated data, respectively. FDR is the expected value of the false discovery proportion and can be computed empirically as

$$\widehat{FDR} = \frac{1}{1000} \sum_{k=1}^{1000} \frac{V_k}{R_k} I(R_k > 0)$$

In our simulations, the decision rule is to reject the null H_{0j} if $fdr(j) = \hat{\pi}_0 \hat{f}_0(j) / \hat{f}(j) < \alpha$. Throughout the simulations, we consider $\alpha = 0.05$. The True Positive Rate, \widehat{TPR} is computed empirically as

$$\widehat{TPR} = \frac{1}{1000} \sum_{k=1}^{1000} \left(\frac{S_k}{S_k + T_k} \right)$$

where S_k and T_k are the number of correctly rejected hypotheses (true discoveries) and the number of falsely accepted hypotheses (false non-discoveries) in the k th generated data, respectively. Three procedures are compared in terms of controlling \widehat{FDR} and \widehat{TPR} , namely the one-stage local FDR procedure, the proposed two-stage procedure and Storey's procedure.

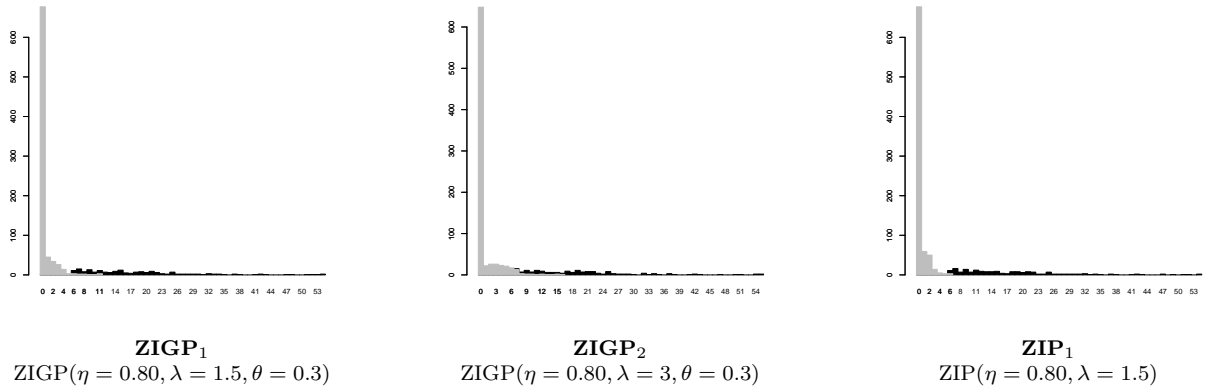


Figure 1. Histogram when the Non-null Distribution is Geometric($p = 0.08$) and $\pi_0 = 0.80$. ZIP₁ represents the well-separated case, ZIGP₁ is the moderately mixed case and ZIGP₂ is the heavily mixed case.

As displayed in Figure 1, the non-null distribution specified is Geometric($p = 0.08$), $\pi_0 = 0.80$ and the fraction of zeros is 0.80. The degree to which the null model is mixed with the non-null model is described using the three cases: ZIP₁, ZIGP₁ and ZIGP₂. The corresponding numerical comparison is shown in Table 1.

Table 1. Numerical Comparison when the Non-null Distribution is Geometric($p = 0.08$), $\pi_0 = 0.80$ and $\alpha = 0.05$. The number in (·) represents the standard deviation.

True f_0	Choice of C	Model for f_0	Two-Stage Procedure			One-Stage Procedure			Storey's FDR		
			R	\widehat{FDR}	\widehat{TPR}	R	\widehat{FDR}	\widehat{TPR}	R	\widehat{FDR}	\widehat{TPR}
ZIGP ₁	C_1	ZIGP	200.86 (21.34)	0.04422 (0.0196)	0.95935 (0.0749)	186.06 (20.13)	0.02992 (0.0150)	0.90634 (0.0675)	175.45 (17.41)	0.02185 (0.0123)	0.85823 (0.0642)
		ZIP	207.58 (15.92)	0.05193 (0.0217)	0.98367 (0.0295)	207.34 (16.31)	0.05176 (0.0219)	0.98297 (0.0311)	197.86 (17.64)	0.04102 (0.0203)	0.94826 (0.0478)
		GP	1.05 (0.22)	0.00000 (0.0000)	0.00525 (0.0011)	0.00 (0.00)	0.0000 (0.0000)	0.00445 (0.0021)	0.00 (0.00)	0.0000 (0.0000)	0.00000 (0.0000)
		P	278.88 (19.55)	0.28142 (0.0453)	1.00000 (0.0000)	278.88 (19.55)	0.28142 (0.0453)	1.00000 (0.0000)	253.81 (14.06)	0.21228 (0.0266)	1.00000 (0.0000)
	C_2	ZIGP	194.94 (28.90)	0.04008 (0.01999)	0.93477 (0.11895)	180.79 (26.94)	0.02717 (0.01525)	0.88500 (0.10929)	170.98 (23.51)	0.02003 (0.01244)	0.83791 (0.10080)
		ZIP	206.71 (16.21)	0.05095 (0.02195)	0.98051 (0.03128)	205.39 (18.18)	0.05010 (0.02288)	0.97600 (0.04315)	196.09 (19.01)	0.03983 (0.02086)	0.94080 (0.05724)
		GP	1.05 (0.22)	0.00000 (0.00000)	0.00525 (0.00112)	0.00 (0.00)	0.00000 (0.00000)	0.00437 (0.00214)	0.00 (0.00)	0.00000 (0.00000)	0.00000 (0.00000)
		P	278.88 (19.55)	0.28142 (0.04527)	1.00000 (0.00000)	278.88 (19.55)	0.28142 (0.04527)	1.00000 (0.00000)	253.81 (14.06)	0.21228 (0.02655)	1.00000 (0.00000)
	C_1	ZIGP	118.39 (75.89)	0.04694 (0.0419)	0.55160 (0.3461)	117.97 (75.64)	0.04646 (0.0417)	0.55160 (0.3461)	115.19 (64.76)	0.03739 (0.0335)	0.54635 (0.2992)
		ZIP	226.68 (19.80)	0.15855 (0.0359)	0.95215 (0.0327)	211.15 (31.92)	0.13399 (0.0529)	0.91567 (0.0708)	191.62 (26.76)	0.10364 (0.0421)	0.85484 (0.07267)
		GP	1.05 (0.23)	0.00000 (0.0000)	0.00528 (0.0012)	0.00 (0.00)	0.0000 (0.0000)	0.00519 (0.0013)	0.00 (0.00)	0.0000 (0.0000)	0.00000 (0.0000)
		P	334.22 (15.36)	0.40181 (0.0269)	1.00000 (0.0000)	334.22 (15.36)	0.40181 (0.0269)	1.00000 (0.0000)	316.02 (17.55)	0.36681 (0.0335)	1.00000 (0.00000)
	C_2	ZIGP	99.80 (78.95)	0.03745 (0.04110)	0.46790 (0.36298)	99.45 (78.65)	0.03701 (0.04082)	0.46784 (0.36290)	99.06 (67.29)	0.02964 (0.03273)	0.47280 (0.31362)
		ZIP	226.45 (19.66)	0.15823 (0.03582)	0.95158 (0.03253)	204.80 (38.10)	0.12532 (0.05961)	0.89378 (0.09831)	187.17 (30.94)	0.09803 (0.04620)	0.83841 (0.09147)
		GP	1.05 (0.23)	0.00000 (0.00000)	0.00528 (0.00119)	0.00 (0.00)	0.00000 (0.00000)	0.00518 (0.00132)	0.00 (0.00)	0.00000 (0.00000)	0.00000 (0.00000)
		P	334.22 (15.36)	0.40181 (0.02692)	1.00000 (0.00000)	334.22 (15.36)	0.40181 (0.02692)	1.00000 (0.00000)	316.02 (17.55)	0.36681 (0.03349)	1.00000 (0.00000)
	C_1	ZIGP	200.85 (13.09)	0.00453 (0.0068)	1.00000 (0.0000)	187.43 (14.19)	0.00095 (0.0023)	0.93620 (0.0260)	182.76 (13.87)	0.00060 (0.0018)	0.91352 (0.0356)
		ZIP	198.14 (14.47)	0.00293 (0.0040)	0.98782 (0.0243)	198.13 (14.48)	0.00293 (0.0040)	0.98782 (0.0243)	191.24 (15.19)	0.00164 (0.0032)	0.95467 (0.0371)
		GP	1.04 (0.22)	0.00000 (0.0000)	0.00525 (0.0011)	0.00 (0.00)	0.0000 (0.0000)	0.00321 (0.0026)	0.00 (0.00)	0.0000 (0.0000)	0.0000 (0.0000)
		P	249.50 (24.88)	0.19300 (0.0731)	1.00000 (0.0000)	249.50 (24.88)	0.19300 (0.0731)	1.0000 (0.0000)	230.52 (13.58)	0.1327 (0.0232)	1.0000 (0.0000)
	C_2	ZIGP	199.39 (13.90)	0.00453 (0.00632)	0.99268 (0.02290)	182.37 (14.89)	0.00061 (0.00182)	0.91191 (0.03567)	174.91 (13.70)	0.00029 (0.00123)	0.87453 (0.03660)
		ZIP	195.07 (15.06)	0.00234 (0.00370)	0.97312 (0.03339)	194.28 (15.81)	0.00229 (0.00370)	0.96964 (0.03918)	188.18 (16.14)	0.00146 (0.00305)	0.93966 (0.04872)
		GP	1.04 (0.22)	0.00000 (0.00000)	0.00525 (0.00113)	0.00 (0.00)	0.00000 (0.00000)	0.00312 (0.00262)	0.00 (0.00)	0.00000 (0.00000)	0.00000 (0.00000)
		P	249.50 (24.88)	0.19300 (0.07306)	1.00000 (0.00000)	249.50 (24.88)	0.19300 (0.07306)	1.00000 (0.00000)	230.52 (13.58)	0.13270 (0.02318)	1.00000 (0.00000)
ZIP ₁	C_1	ZIGP	200.85 (13.09)	0.00453 (0.0068)	1.00000 (0.0000)	187.43 (14.19)	0.00095 (0.0023)	0.93620 (0.0260)	182.76 (13.87)	0.00060 (0.0018)	0.91352 (0.0356)
		ZIP	198.14 (14.47)	0.00293 (0.0040)	0.98782 (0.0243)	198.13 (14.48)	0.00293 (0.0040)	0.98782 (0.0243)	191.24 (15.19)	0.00164 (0.0032)	0.95467 (0.0371)
		GP	1.04 (0.22)	0.00000 (0.0000)	0.00525 (0.0011)	0.00 (0.00)	0.0000 (0.0000)	0.00321 (0.0026)	0.00 (0.00)	0.0000 (0.0000)	0.0000 (0.0000)
		P	249.50 (24.88)	0.19300 (0.0731)	1.00000 (0.0000)	249.50 (24.88)	0.19300 (0.0731)	1.0000 (0.0000)	230.52 (13.58)	0.1327 (0.0232)	1.0000 (0.0000)
	C_2	ZIGP	199.39 (13.90)	0.00453 (0.00632)	0.99268 (0.02290)	182.37 (14.89)	0.00061 (0.00182)	0.91191 (0.03567)	174.91 (13.70)	0.00029 (0.00123)	0.87453 (0.03660)
		ZIP	195.07 (15.06)	0.00234 (0.00370)	0.97312 (0.03339)	194.28 (15.81)	0.00229 (0.00370)	0.96964 (0.03918)	188.18 (16.14)	0.00146 (0.00305)	0.93966 (0.04872)
		GP	1.04 (0.22)	0.00000 (0.00000)	0.00525 (0.00113)	0.00 (0.00)	0.00000 (0.00000)	0.00312 (0.00262)	0.00 (0.00)	0.00000 (0.00000)	0.00000 (0.00000)
		P	249.50 (24.88)	0.19300 (0.07306)	1.00000 (0.00000)	249.50 (24.88)	0.19300 (0.07306)	1.00000 (0.00000)	230.52 (13.58)	0.13270 (0.02318)	1.00000 (0.00000)

Overall, there are more rejections using C_1 as a cut-off compared to C_2 . This suggests that the extension of Efron’s method is conservative and would miss significant positions. Also, even if using C_1 yields more rejections, it still controls the value of FDR indicating the superiority of C_1 as a cut-off method.

The difference between C_1 and C_2 is further highlighted for ZIGP₂, where the true null distribution is heavily mixed with the non-null distribution and overdispersion is also present. When the model used for the estimation of f_0 is ZIGP, the value of \widehat{TPR} is relatively higher using C_1 , while keeping the \widehat{FDR} controlled.

The results for the three null models can also be compared. Since null and non-null distribution is moderately mixed for ZIGP₁, the resulting \widehat{TPR} for all three procedures is substantially higher than the \widehat{TPR} for ZIGP₂, regardless of the model used for the estimation of f_0 . Given that \widehat{FDR} is controlled in all procedures if the model for f_0 is ZIGP, the Two-Stage procedure yields the highest \widehat{TPR} compared to the One-Stage local FDR and Storey’s procedure. This suggests that the proposed procedure is better than the other existing procedures.

Meanwhile, ZIGP₁ is allowed to vary from ZIP₁ in terms of the overdispersion parameter θ . Due to the “well-separation” if the true null is ZIP₁, then the \widehat{TPR} for ZIP₁ is slightly higher than the \widehat{TPR} for ZIGP₁. Moreover, the \widehat{FDR} for all three procedures for ZIP₁ are noticeably lower than the \widehat{FDR} for ZIGP₁. This means that the number of rejections for ZIGP₁ and ZIP₁ are almost the same but there are more erroneous rejections for ZIGP₁. This result can be explained by the presence of overdispersion in ZIGP₁ as compared to ZIP₁.

Figure 2 presents the histograms when the non-null distribution specified is Binomial, the proportion of null cases is 0.80 and the fraction of zeros is 0.40. Unlike the parametrization of the Geometric non-null distribution which appears to be skewed to the right, this non-null distribution exhibits near symmetry. Similar to the previous set of results, the true null distribution is allowed to vary in terms of λ and θ . In terms of the mixing of the null and non-null distribution, ZIP₂ represents the well-separated case, ZIGP₃ is the moderately mixed case while ZIGP₄ can be described as the heavily mixed case. The respective numerical comparison is shown in Table 2.

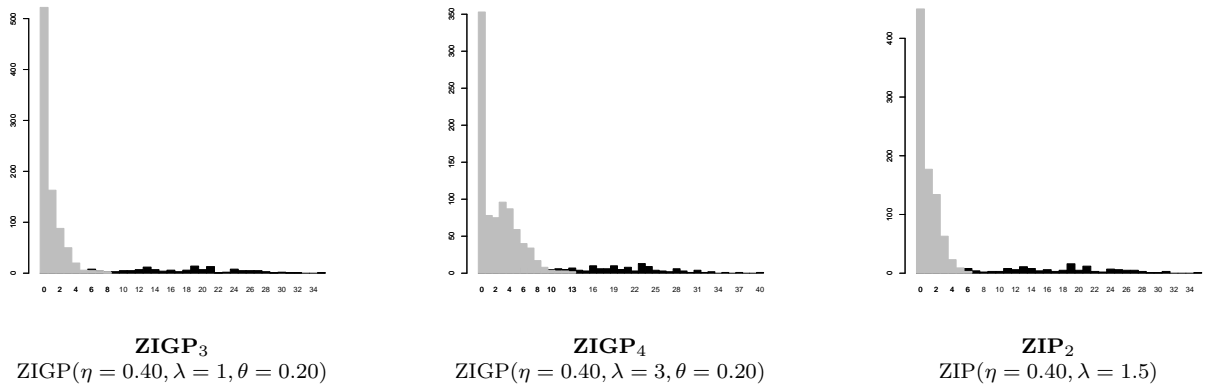


Figure 2. Histogram when the Non-null Distribution is Binomial($n = 250, p = 0.20$) and $\pi_0 = 0.80$. ZIP₂ represents the well-separated case, ZIGP₃ is the moderately mixed case and ZIGP₄ is the heavily mixed case.

Table 2. Numerical Comparison when the Non-null Distribution is Binomial($n = 250, p = 0.20$), $\pi_0 = 0.80$ and $\alpha = 0.05$. The number in (·) represents the standard deviation.

True f_0	Choice of C	Model for f_0	Two-Stage Procedure			One-Stage Procedure			Storey's FDR		
			R	\widehat{FDR}	\widehat{TPR}	R	\widehat{FDR}	\widehat{TPR}	R	\widehat{FDR}	\widehat{TPR}
ZIGP ₃	C_1	ZIGP	209.45 (13.76)	0.04586 (0.02707)	1.00000 (0.00000)	186.45 (12.64)	0.00647 (0.00646)	0.92738 (0.01928)	190.90 (12.81)	0.01274 (0.00886)	0.94354 (0.01817)
		ZIP	203.71 (14.31)	0.02861 (0.01380)	0.99039 (0.01957)	203.57 (14.45)	0.02835 (0.01405)	0.99039 (0.01957)	203.38 (14.35)	0.02848 (0.01405)	0.98897 (0.02258)
		GP	68.03 (80.14)	0.00083 (0.00412)	0.34534 (0.40543)	65.57 (78.36)	0.00039 (0.00166)	0.33245 (0.39491)	72.13 (81.72)	0.00044 (0.00197)	0.36634 (0.41397)
		P	255.54 (25.60)	0.21280 (0.07095)	1.00000 (0.00000)	255.54 (25.60)	0.21280 (0.07095)	1.00000 (0.00000)	234.19 (13.84)	0.14702 (0.02419)	1.00000 (0.00000)
	C_2	ZIGP	209.68 (13.76)	0.04703 (0.02436)	1.00000 (0.00000)	185.32 (12.75)	0.00550 (0.00619)	0.92263 (0.02064)	188.87 (12.94)	0.00990 (0.00809)	0.93613 (0.02037)
		ZIP	203.79 (14.15)	0.02875 (0.01382)	0.99069 (0.01959)	203.65 (14.30)	0.02849 (0.01407)	0.99069 (0.01959)	203.68 (14.17)	0.02888 (0.01406)	0.99004 (0.02183)
		GP	64.52 (79.42)	0.00082 (0.00412)	0.32815 (0.40250)	62.08 (77.63)	0.00041 (0.00170)	0.31511 (0.39230)	67.48 (81.83)	0.00055 (0.00219)	0.34355 (0.41528)
		P	255.54 (25.60)	0.21280 (0.07095)	1.00000 (0.00000)	255.54 (25.60)	0.21280 (0.07095)	1.00000 (0.00000)	234.19 (13.84)	0.14702 (0.02419)	1.00000 (0.00000)
ZIGP ₄	C_1	ZIGP	174.26 (43.43)	0.04521 (0.03435)	0.82936 (0.19135)	160.76 (41.06)	0.02237 (0.02015)	0.82550 (0.19064)	173.01 (33.78)	0.03546 (0.02505)	0.83355 (0.14681)
		ZIP	254.22 (15.19)	0.24384 (0.02873)	0.96212 (0.01405)	205.90 (26.80)	0.10126 (0.06698)	0.93342 (0.02812)	206.28 (19.78)	0.10221 (0.04791)	0.92392 (0.02593)
		GP	1.20 (0.53)	0.00000 (0.00000)	0.00604 (0.00265)	0.00 (0.00)	0.00000 (0.00000)	0.00604 (0.00265)	13.22 (6.14)	0.00000 (0.00000)	0.06649 (0.03121)
		P	538.18 (41.90)	0.62710 (0.03245)	1.00000 (0.00000)	538.18 (41.90)	0.62710 (0.03245)	1.00000 (0.00000)	476.05 (47.22)	0.57682 (0.04521)	1.00000 (0.00000)
	C_2	ZIGP	98.53 (89.67)	0.02540 (0.03596)	0.47057 (0.42354)	89.99 (84.25)	0.01500 (0.02083)	0.46826 (0.42133)	115.13 (69.14)	0.02001 (0.02697)	0.55907 (0.32719)
		ZIP	254.22 (15.19)	0.24384 (0.02873)	0.96212 (0.01405)	204.28 (30.40)	0.10051 (0.07169)	0.92299 (0.05145)	205.19 (22.79)	0.10092 (0.05353)	0.91887 (0.03732)
		GP	1.20 (0.53)	0.00000 (0.00000)	0.00604 (0.00265)	0.00 (0.00)	0.00000 (0.00000)	0.00604 (0.00265)	8.34 (3.35)	0.00000 (0.00000)	0.04179 (0.01653)
		P	538.18 (41.90)	0.62710 (0.03245)	1.00000 (0.00000)	538.18 (41.90)	0.62710 (0.03245)	1.00000 (0.00000)	476.05 (47.22)	0.57682 (0.04521)	1.00000 (0.00000)
ZIP ₂	C_1	ZIGP	201.82 (12.93)	0.01046 (0.00823)	0.99983 (0.00170)	184.85 (12.68)	0.00028 (0.00126)	0.92513 (0.02015)	187.86 (12.85)	0.00110 (0.00270)	0.93943 (0.01981)
		ZIP	190.88 (13.01)	0.00261 (0.00456)	0.95307 (0.01783)	188.15 (13.32)	0.00156 (0.00421)	0.95103 (0.01763)	190.92 (12.92)	0.00266 (0.00449)	0.95327 (0.01957)
		GP	195.38 (15.79)	0.00689 (0.00790)	0.97134 (0.04563)	179.43 (13.93)	0.00015 (0.00090)	0.89725 (0.04024)	184.31 (13.40)	0.00067 (0.00221)	0.92223 (0.03359)
		P	244.13 (28.99)	0.17402 (0.07862)	1.00000 (0.00000)	244.13 (28.99)	0.17402 (0.07862)	1.00000 (0.00000)	222.66 (17.54)	0.10062 (0.04938)	1.00000 (0.00000)
	C_2	ZIGP	201.55 (13.13)	0.01087 (0.00922)	0.99807 (0.01027)	181.79 (12.63)	0.00010 (0.00073)	0.90989 (0.02130)	186.02 (12.50)	0.00038 (0.00147)	0.93100 (0.01900)
		ZIP	190.93 (13.01)	0.00263 (0.00464)	0.95328 (0.01796)	187.68 (13.40)	0.00147 (0.00420)	0.94782 (0.01967)	190.62 (13.11)	0.00259 (0.00465)	0.95181 (0.02136)
		GP	195.19 (18.21)	0.00714 (0.00785)	0.97024 (0.06445)	178.59 (16.38)	0.00015 (0.00091)	0.89298 (0.05861)	183.63 (15.85)	0.00066 (0.00220)	0.91884 (0.05344)
		P	244.13 (28.99)	0.17402 (0.07862)	1.00000 (0.00000)	244.13 (28.99)	0.17402 (0.07862)	1.00000 (0.00000)	222.66 (17.54)	0.10062 (0.04938)	1.00000 (0.00000)

The difference between C_1 and C_2 is apparent for ZIGP_4 , where there is overdispersion and the true null distribution is heavily mixed with the non-null distribution. If ZIGP is the model used for the estimation of f_0 , the value of \widehat{TPR} is substantially higher using C_1 , while keeping the \widehat{FDR} controlled.

According to Table 2, the resulting \widehat{TPR} for ZIGP_3 is substantially higher than the \widehat{TPR} for ZIGP_4 , regardless of the model used for the estimation of f_0 and the procedure employed. Given that \widehat{FDR} is controlled in all procedures for ZIGP_3 if the model used for the estimation of f_0 is ZIGP, this suggests that the Two-Stage procedure is better than the One-Stage procedure and Storey's procedure. However, for the scenario specified in ZIGP_4 , the Storey's procedure is slightly better than the Two-Stage procedure if ZIGP is the model used for f_0 .

It can also be noted that for ZIP_2 , the number of erroneous rejections is lesser if the model used for the estimation of f_0 is ZIP as compared to ZIGP. However, given that \widehat{FDR} is controlled by specifying either of the two models, using ZIGP leads to a higher \widehat{TPR} than when the true model ZIP is specified. This result implies using ZIGP would yield satisfactory results even under model misspecification.

As presented in Figure 3, the non-null distribution considered is also Binomial, fraction of zeros is still 0.40 but the proportion of null cases is reduced to 0.35. Again, the true null distribution is allowed to vary in terms of λ and θ . ZIP_3 represents the well-separated case, ZIGP_5 is the moderately mixed case while ZIGP_6 can be described as the overdispersed and heavily mixed case. The respective numerical comparison is shown in Table 3.

Based on the results shown in Table 3, using C_1 as a cut-off resulted to more rejections in the case of ZIGP_5 and ZIP_3 . However, contrary to the results from Table 1 and 2, there are more rejections using C_2 for ZIGP_6 , where there is overdispersion and the true null distribution is heavily mixed with the non-null distribution. If ZIGP is the model used for the estimation of f_0 , the value of \widehat{TPR} is substantially higher using Storey's procedure, while keeping the \widehat{FDR} controlled.

Furthermore, the resulting \widehat{TPR} for ZIGP_5 is substantially higher than the \widehat{TPR} for ZIGP_6 , regardless of the model used for the estimation of f_0 and the procedure employed.

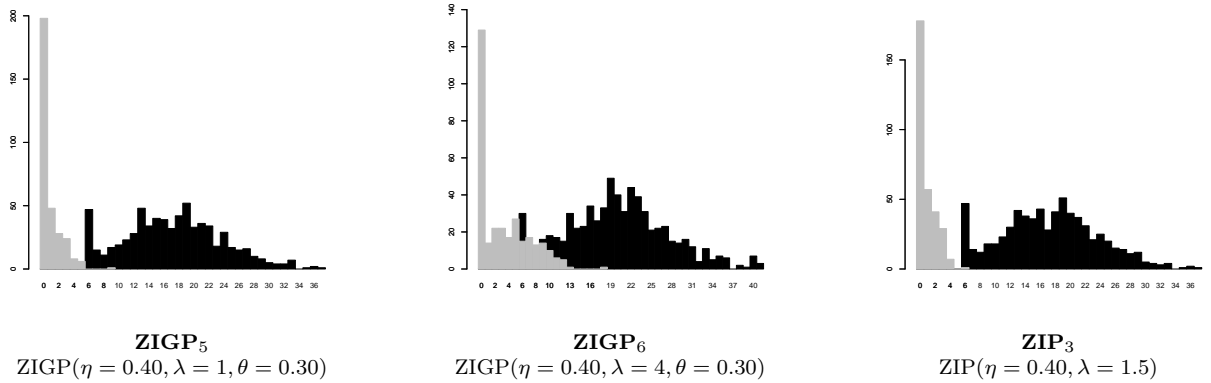


Figure 3. Histogram when the Non-null Distribution is Binomial($n = 250, p = 0.20$) and $\pi_0 = 0.35$. ZIP_3 represents the well-separated case, ZIGP_5 is the moderately mixed case and ZIGP_6 is the heavily mixed case.

Table 3. Numerical Comparison when the Non-null Distribution is Binomial($n = 250, p = 0.20$), $\pi_0 = 0.35$ and $\alpha = 0.05$. The number in (\cdot) represents the standard deviation.

True f_0	Choice of C	Model for f_0	Two-Stage Procedure			One-Stage Procedure			Storey's FDR		
			R	\widehat{FDR}	\widehat{TPR}	R	\widehat{FDR}	\widehat{TPR}	R	\widehat{FDR}	\widehat{TPR}
ZIGP ₅	C_1	ZIGP	655.70 (15.99)	0.01105 (0.00543)	0.99796 (0.01035)	645.10 (24.24)	0.00886 (0.00450)	0.98449 (0.02442)	621.88 (18.86)	0.00620 (0.00355)	0.95109 (0.01700)
		ZIP	656.25 (15.15)	0.01040 (0.00434)	0.99945 (0.00492)	656.21 (15.22)	0.01039 (0.00435)	0.99945 (0.00492)	643.14 (23.35)	0.00894 (0.00462)	0.98086 (0.02551)
		GP	1.24 (0.59)	0.00000 (0.00000)	0.00191 (0.00090)	0.00 (0.00)	0.00000 (0.00000)	0.00191 (0.00090)	64.03 (23.57)	0.00000 (0.00000)	0.09831 (0.03550)
		P	676.98 (14.52)	0.04008 (0.01562)	1.00000 (0.00000)	676.98 (14.52)	0.04008 (0.01562)	1.00000 (0.00000)	672.19 (15.21)	0.03327 (0.01160)	1.00000 (0.00000)
	C_2	ZIGP	643.89 (80.65)	0.01296 (0.00675)	0.97781 (0.11962)	634.14 (80.75)	0.00884 (0.00465)	0.96730 (0.11952)	612.78 (56.92)	0.00605 (0.00369)	0.93724 (0.08339)
		ZIP	657.03 (15.49)	0.01151 (0.00577)	0.99950 (0.00512)	656.97 (15.66)	0.01150 (0.00578)	0.99950 (0.00512)	654.92 (19.28)	0.01251 (0.00647)	0.99522 (0.01547)
		GP	1.24 (0.59)	0.00000 (0.00000)	0.00191 (0.00090)	0.00 (0.00)	0.00000 (0.00000)	0.00191 (0.00090)	67.67 (25.38)	0.00000 (0.00000)	0.10389 (0.03816)
		P	681.10 (13.56)	0.04588 (0.01729)	1.00000 (0.00000)	681.10 (13.56)	0.04588 (0.01729)	1.00000 (0.00000)	673.06 (14.52)	0.03455 (0.01102)	1.00000 (0.00000)
	C_1	ZIGP	307.49 (260.59)	0.01984 (0.02601)	0.45465 (0.37898)	307.47 (260.61)	0.01990 (0.02603)	0.45465 (0.37898)	383.81 (189.41)	0.01954 (0.02387)	0.57293 (0.27060)
		ZIP	702.05 (26.90)	0.10970 (0.02259)	0.96127 (0.01592)	592.67 (38.38)	0.04123 (0.01975)	0.87349 (0.04190)	594.67 (33.72)	0.04115 (0.01704)	0.87694 (0.03544)
		GP	1.37 (2.81)	0.00002 (0.00074)	0.00209 (0.00416)	0.19 (2.77)	0.00488 (0.01091)	0.00209 (0.00416)	92.38 (27.92)	0.00040 (0.00212)	0.14189 (0.04198)
		P	823.15 (12.27)	0.21052 (0.01884)	1.00000 (0.00000)	823.15 (12.27)	0.21052 (0.01884)	1.00000 (0.00000)	805.39 (15.15)	0.19304 (0.01961)	1.00000 (0.00000)
	C_2	ZIGP	319.34 (217.41)	0.01029 (0.01668)	0.48100 (0.32111)	319.31 (217.44)	0.01029 (0.01668)	0.48100 (0.32111)	390.02 (159.49)	0.01116 (0.01647)	0.58950 (0.23071)
		ZIP	712.53 (25.11)	0.11166 (0.01854)	0.97368 (0.01448)	665.87 (56.82)	0.07683 (0.03946)	0.94700 (0.04093)	653.12 (44.11)	0.06752 (0.03039)	0.93548 (0.03306)
		GP	1.31 (2.00)	0.00000 (0.00000)	0.00201 (0.00300)	0.11 (1.94)	0.00000 (0.00000)	0.00201 (0.00300)	95.50 (27.36)	0.00008 (0.00095)	0.14667 (0.04083)
		P	823.43 (12.85)	0.21082 (0.01674)	1.00000 (0.00000)	823.43 (12.85)	0.21082 (0.01674)	1.00000 (0.00000)	803.25 (15.06)	0.19096 (0.01679)	1.00000 (0.00000)
	C_1	ZIGP	650.68 (14.59)	0.00139 (0.00150)	1.00000 (0.00000)	636.68 (23.85)	0.00092 (0.00137)	0.97882 (0.02500)	619.64 (17.52)	0.00038 (0.00086)	0.95323 (0.01465)
		ZIP	650.36 (14.88)	0.00139 (0.00150)	0.99952 (0.00448)	650.35 (14.91)	0.00139 (0.00150)	0.99952 (0.00448)	636.75 (22.60)	0.00094 (0.00137)	0.97899 (0.02522)
		GP	107.04 (239.05)	0.00026 (0.00091)	0.16526 (0.36896)	105.11 (237.61)	0.00138 (0.00171)	0.16393 (0.36605)	184.21 (195.71)	0.00011 (0.00055)	0.28388 (0.30250)
		P	655.00 (15.64)	0.00783 (0.01538)	1.00000 (0.00000)	655.00 (15.64)	0.00783 (0.01538)	1.00000 (0.00000)	654.29 (14.22)	0.00689 (0.00593)	1.00000 (0.00000)
	C_2	ZIGP	641.71 (72.01)	0.00208 (0.00256)	0.98573 (0.10856)	610.09 (70.63)	0.00037 (0.00094)	0.93888 (0.10539)	610.13 (48.38)	0.00026 (0.00069)	0.93889 (0.07204)
		ZIP	644.72 (20.33)	0.00120 (0.00147)	0.99096 (0.01905)	643.76 (21.87)	0.00118 (0.00147)	0.99085 (0.01932)	634.46 (23.03)	0.00099 (0.00164)	0.97541 (0.02611)
		GP	106.40 (238.38)	0.00026 (0.00090)	0.16431 (0.36802)	104.47 (236.93)	0.00136 (0.00170)	0.16299 (0.36511)	186.78 (193.93)	0.00011 (0.00055)	0.28780 (0.29983)
		P	660.72 (14.82)	0.01636 (0.02008)	1.00000 (0.00000)	660.72 (14.82)	0.01636 (0.02008)	1.00000 (0.00000)	657.67 (12.48)	0.01204 (0.00911)	1.00000 (0.00000)
ZIP ₃	C_1	ZIGP	650.68 (14.59)	0.00139 (0.00150)	1.00000 (0.00000)	636.68 (23.85)	0.00092 (0.00137)	0.97882 (0.02500)	619.64 (17.52)	0.00038 (0.00086)	0.95323 (0.01465)
		ZIP	650.36 (14.88)	0.00139 (0.00150)	0.99952 (0.00448)	650.35 (14.91)	0.00139 (0.00150)	0.99952 (0.00448)	636.75 (22.60)	0.00094 (0.00137)	0.97899 (0.02522)
		GP	107.04 (239.05)	0.00026 (0.00091)	0.16526 (0.36896)	105.11 (237.61)	0.00138 (0.00171)	0.16393 (0.36605)	184.21 (195.71)	0.00011 (0.00055)	0.28388 (0.30250)
		P	655.00 (15.64)	0.00783 (0.01538)	1.00000 (0.00000)	655.00 (15.64)	0.00783 (0.01538)	1.00000 (0.00000)	654.29 (14.22)	0.00689 (0.00593)	1.00000 (0.00000)
	C_2	ZIGP	641.71 (72.01)	0.00208 (0.00256)	0.98573 (0.10856)	610.09 (70.63)	0.00037 (0.00094)	0.93888 (0.10539)	610.13 (48.38)	0.00026 (0.00069)	0.93889 (0.07204)
		ZIP	644.72 (20.33)	0.00120 (0.00147)	0.99096 (0.01905)	643.76 (21.87)	0.00118 (0.00147)	0.99085 (0.01932)	634.46 (23.03)	0.00099 (0.00164)	0.97541 (0.02611)
		GP	106.40 (238.38)	0.00026 (0.00090)	0.16431 (0.36802)	104.47 (236.93)	0.00136 (0.00170)	0.16299 (0.36511)	186.78 (193.93)	0.00011 (0.00055)	0.28780 (0.29983)
		P	660.72 (14.82)	0.01636 (0.02008)	1.00000 (0.00000)	660.72 (14.82)	0.01636 (0.02008)	1.00000 (0.00000)	657.67 (12.48)	0.01204 (0.00911)	1.00000 (0.00000)

Another scenario considered is when the true non-null distribution is Geometric, the proportion of null cases is 0.85 but the fraction of zeros is 0.40. Unlike the scenario presented in Table 1 and Figure 1, this means that the specified proportion of zeros is reduced to half. The interest is to determine whether there would be a change in pattern should there be a significant decrease in the number of positions without a mutation. The histograms are displayed in Figure 4 and the corresponding numerical comparisons are presented in Tables 4 and 5 found in the Supplementary section. It can be noted that regardless of the magnitude of the fraction of zeros, a similar pattern can be observed in terms of the superiority of C_1 as a method for choosing C . However, for the heavily mixed case presented in ZIGP₈, the \widehat{FDR} for the two-stage procedure is slightly higher than the specified level which is 0.05.

In addition, another scenario considered is when the true non-null distribution is Binomial($n = 250, p = 0.20$), $\pi_0 = 0.70$ and η is 0.40. The goal is to determine whether there would be a change in pattern of results if there is a drop in the proportion of the null cases in the mixture model as compared to the results in Table 2 and Figure 2. The histograms are displayed in Figure 5 and the numerical comparisons are presented in Tables 6 and 7 also found in the Supplementary section. Results revealed that even with the decrease in the value of π_0 , a similar pattern can be observed in terms of the superiority of C_1 as a method for choosing C particularly for the overdispersed and heavily mixed case presented in ZIGP₁₀. Moreover, for ZIGP₁₀, Storey's procedure yielded more rejections and a higher \widehat{TPR} compared to the local FDR procedure where one-stage and two-stage procedure results coincided.

Overall, for the well-separated and moderately mixed case, if the null model is correctly specified, using the Two-Stage procedure yields \widehat{FDR} closest to the nominal level α . Consequently, the Two-Stage procedure is superior in terms of \widehat{TPR} in most cases. If the true null model is ZIGP and the null model is correctly specified, \widehat{FDR} is controlled in all procedures. However, the Two-Stage procedure is better than the One-Stage procedure and Storey's procedure in terms of \widehat{TPR} .

It can also be noted that if the true model is ZIP and ZIGP is used to model the null distribution, then the Two-Stage Procedure still yields the closest \widehat{FDR} to α and leads to higher \widehat{TPR} as compared to the other procedures. This implies using the Two-Stage Procedure when the null model is misspecified would still produce satisfactory results. Moreover, regardless of the shape of the non-null distribution, the Two-Stage Procedure yields better results than the other procedures.

4.2 Application to Protein Domain Data

One interesting issue is identifying the position of somatic mutations, so called hotspot, on protein domains. The key question is among fixed number of positions in a single domain, which ones are significantly different from the majority. It is a novel solution for the identification of driver mutations which lead tumor progression in somatic tumor samples and recapitulates much of what is known about how protein domain families contribute to the initiation or progression of cancer.

As an example, we analyze the mutation data which were obtained from the tumors of 5,848 patients from The Cancer Genome Atlas (TCGA) data portal (<http://tcga-data.nci.nih.gov/tcga/>, Collins and Barker, 2007). These were mapped to specific positions within protein domain models to identify clusters. TCGA MAF files were obtained on July 7th, 2014 for 20 cancer types: Adrenocortical Carcinoma (ACC), Bladder Urothelial Carcinoma (BLCA), Brain Lower Grade Glioma (LGG), Breast Invasive Carcinoma (BRCA), Colon Adenocarcinoma (COAD), Glioblastoma Multiforme (GBM), Head and Neck Squamous Cell Carcinoma (HNSC), Kidney Chromophobe (KICH), Kidney Renal Clear Cell Carcinoma (KIRC), Liver Hepatocellular Carcinoma (LIHC), Lung Adenocarcinoma (LUAD), Lung Squamous Cell Carcinoma (LUSC), Ovarian Serous Cystadenocarcinoma (OV), Pancreatic Adenocarcinoma (PAAD), Prostate Adenocarcinoma (PRAD), Rectum Adenocarcinoma (READ), Skin Cutaneous Melanoma (SKCM), Stomach Adenocarcinoma (STAD), Thyroid Carcinoma (THCA), and Uterine Corpus Endometrial Carcinoma (UCEC). The mutations were mapped to proteins and domain models (Peterson et al., 2010 and Peterson et al., 2012).

Among several hundreds of domains, we focus on five functionally well-known domains to identify the hotspots in TCGA/GBF dataset. We start with the hotspots on growth factors (cd00031), which are known to harbor reoccurring somatic mutations involved with clonal expansion, invasion across tissue barriers, and colonization of distant niches ([19]; [48]; [51]). Furthermore, protein kinases (cd00180) and the RAS-Like GTPase family of genes (cd00882), which are well-known for their role in regulating pathways important to cancer ([1]; [7]; [3]; [33]; [49]). Genes with kinases or RAS-Like GTPases are expected to harbor driver mutations that reoccur at specific sites since they are classic examples of proto-oncogenes that mutate into oncogenes, contributing to cancer ([2]; [5]). Additionally, we identify hotspots on ankyrin domains (cd00204), which play a role in mediating protein-protein interactions important in cancer ([27]; [18]). Furthermore, we find hotspots on transmembrane domains of proteins that are known to be involved with signal transduction, which is relevant in controlling processes involved with cancer ([41]; [43]) and experimental evidence confirms the important regulatory role played by membrane proteins in cancer ([22]; [26]; [32]; [42]; [30];

Since the mutation counts are discrete, we apply our proposed method based on various discrete models, such as Zero-Inflated Generalized Poisson, Zero-Inflated Poisson, Generalized Poisson and ordinary Poisson distribution for f_0 . The estimated parameters based on those models are reported in Table 8 and the identified number of positions which are mutated differently from expected are in Table 9. Figure 6 shows the distribution of each protein domain and its total number of positions.

For example, when we conduct hypothesis testing framework of section 3 to identify hotspots under the assumption of f_0 follows ZIGP, the results show that the identified hotspots on growth factor domain (cd00031) based on one stage and two procedures are 143 positions based on C_2 among total of 366 positions. On the other hand, the local FDR with C_1 identifies more hotspots for two stage (201) than one stage (191) and Storey's procedure (200). Rest of domains can be analyzed in the similar manner.

Table 8. Comparison of Parameter Estimates for Protein Domain Data

Data	Model for f_0	C_1							C_2						
		η	λ	θ	π	C	D		η	λ	θ	π	C	D	
cd00031	ZIGP	0.3246	1.9168	0.1416	0.4576	6	7		0.2253	2.1449	0.5738	0.6139	4	36	
	ZIP	0.2289	1.0949	NA	0.3985	3	6		0.2760	1.3856	NA	0.4244	2	6	
	GP	NA	1.5559	0.6609	0.5540	11	36		NA	2.0944	0.6668	0.8321	3	36	
	P	NA	0.8082	NA	0.3944	3	6		NA	0.7994	NA	0.3929	2	6	
cd00180	ZIGP	0.5773	1.7754	0.2255	0.7095	7	10		0.4569	2.0310	0.7021	0.8716	5	63	
	ZIP	0.5507	1.1095	NA	0.6588	3	7		0.5331	0.9701	NA	0.6484	2	7	
	GP	NA	1.3097	0.8292	0.8379	17	63		NA	1.5161	0.8287	0.9925	7	63	
	P	NA	0.2419	NA	0.5864	1	7		NA	0.2419	NA	0.5864	1	7	
cd00204	ZIGP	0.5060	1.2628	0.0002	0.6853	5	6		0.5062	1.2645	0.0002	0.6854	4	5	
	ZIP	0.1287	0.5081	NA	0.6784	3	7		0.1403	0.5188	NA	0.6792	2	7	
	GP	NA	1.1923	0.7275	0.7801	12	34		NA	1.2048	0.7372	0.7908	10	34	
	P	NA	0.4409	NA	0.6780	3	7		NA	0.4089	NA	0.6661	1	7	
cd00882	ZIGP	0.6736	1.3969	0.0003	0.8003	4	5		0.6736	1.3969	0.0003	0.8003	4	5	
	ZIP	0.5201	0.6786	NA	0.7907	3	7		0.5136	0.6616	NA	0.7896	2	7	
	GP	NA	1.2716	0.7423	1.0000	9	25		NA	1.2888	0.7425	1.0000	7	25	
	P	NA	0.2174	NA	0.7503	1	7		NA	0.2174	NA	0.7503	1	7	
pfam00001	ZIGP	0.0526	2.4020	0.3839	0.4031	13	18		0.0009	7.5244	0.7562	1.0000	1	233	
	ZIP	0.0000	44.9641	NA	1.0000	18	45		NA	44.9641	NA	1.0000	18	45	
	GP	NA	2.2464	0.4164	0.4048	13	21		NA	4.8034	0.7937	1.0000	2	233	
	P	NA	3.7966	NA	0.4116	19	20		NA	3.7966	NA	0.4116	19	20	

Table 9. Comparison of Number of Rejections for Protein Domain Data

Data	Method	One-Stage Procedure				Two-Stage Procedure				Storey's FDR			
		ZIGP	ZIP	GP	P	ZIGP	ZIP	GP	P	ZIGP	ZIP	GP	P
cd00031	C_1	191	212	140	212	201	212	141	212	200	211	154	211
	C_2	143	205	16	212	143	205	17	212	162	204	85	211
cd00180	C_1	248	288	0	326	251	288	1	326	247	270	5	300
	C_2	63	288	0	326	63	288	1	326	122	284	0	300
cd00204	C_1	129	130	19	130	129	130	20	130	125	128	60	128
	C_2	129	130	12	130	130	130	13	130	125	128	52	128
cd00882	C_1	148	155	0	169	155	155	1	169	147	154	2	160
	C_2	148	155	0	169	155	155	1	169	147	154	2	160
pfam00001	C_1	255	340	253	265	255	405	253	265	242	206	240	254
	C_2	87	340	57	265	87	405	57	265	148	206	174	254

Results from Table 9 revealed that using C_1 yields more rejections. This suggests that the data analysis for the real data shows the same pattern as the simulation results presented previously. Moreover, two domains can be highlighted in terms of the difference in the number of rejections, namely, cd00180 and pfam00001. The number of rejections using C_1 is almost four times higher if the model used for f_0 is ZIGP and the procedure employed is either local FDR or two-stage method. Using Storey's FDR, the number of rejections using C_1 is almost twice given that the model for f_0 is ZIGP. Overall, the results for the real data analysis is consistent with the simulation studies.

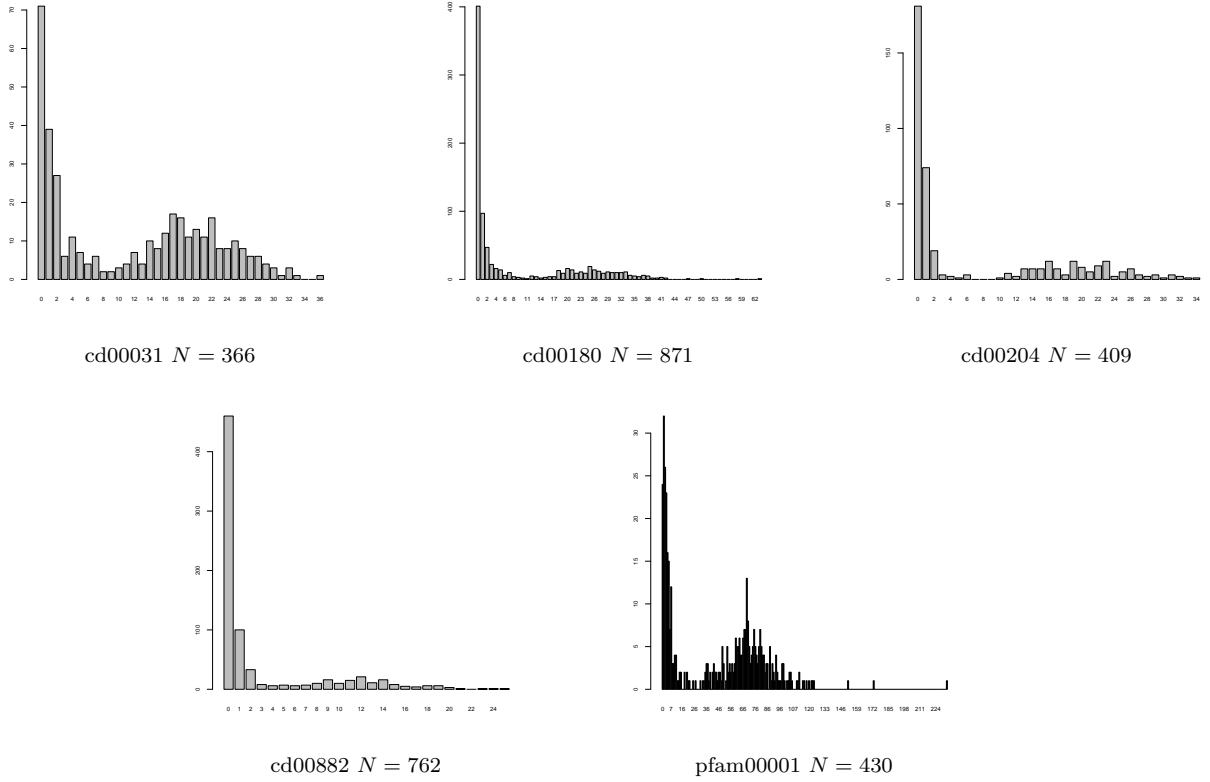


Figure 6. Histogram of Protein Domain Data

5 Conclusion

In this paper, our main interest is to select significant mutation counts while controlling a given level of Type I error via False Discovery Rate (FDR) procedures. We assume that if the number of mutations $a \leq C$, then a is guaranteed to be from the null model, for some positive integer C . We propose a method for identify a cut-off C and show that this is superior to the cut-off developed by extending Efron's proposal. In addition, after the selection of this cut-off, we consider a screening process so that the number of mutations exceeding a certain value D ($D > C$) should be considered as significant mutations. This two-stage procedure in the selection of C and D yielded a testing procedure with increased power compared to Efron's local FDR and Storey's FDR particularly if the non-null distribution behaves similarly to the Geometric distribution and if the null distribution is well-separated and overdispersion is not observed.

References

- [1] Young-Ho Ahn, Yanan Yang, Don L Gibbons, Chad J Creighton, Fei Yang, Ignacio I Wistuba, Wei Lin, Nishan Thilaganathan, Cristina A Alvarez, Jonathon Roybal, et al. Map2k4 functions as a tumor suppressor in lung adenocarcinoma and inhibits tumor cell invasion by decreasing peroxisome proliferator-activated receptor γ 2 expression. *Molecular and cellular biology*, 31(21):4270–4285, 2011.
- [2] Marshall W Anderson, Steven H Reynolds, Ming You, and Robert M Maronpot. Role of proto-oncogene activation in carcinogenesis. *Environmental health perspectives*, 98:13, 1992.
- [3] Katharina Balschun, Jochen Haag, Ann-Kathrin Wenke, Witigo von Schönfels, Nicolas T Schwarz, and Christoph Röcken. Kras, nras, pik3ca exon 20, and braf genotypes in synchronous and metachronous primary colorectal cancers: Diagnostic and therapeutic implications. *The Journal of Molecular Diagnostics*, 13(4):436–445, 2011.
- [4] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.
- [5] Martin J Cline. Keynote address: The role of proto-oncogenes in human cancer: Implications for diagnosis and treatment. *International Journal of Radiation Oncology* Biology* Physics*, 13(9):1297–1301, 1987.
- [6] PC Consul and GC Jain. On the generalization of Poisson distribution. In *Annals of Mathematical Statistics*, volume 41, page 1387, 1970.
- [7] Helen Davies, Chris Hunter, Raffaella Smith, Philip Stephens, Chris Greenman, Graham Bignell, Jon Teague, Adam Butler, Sarah Edkins, Claire Stevens, et al. Somatic mutations of the protein kinase gene family in human lung cancer. *Cancer research*, 65(17):7591–7595, 2005.
- [8] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.
- [9] Sandrine Dudoit, Juliet Popper Shaffer, and Jennifer C Boldrick. Multiple hypothesis testing in microarray experiments. *Statistical Science*, pages 71–103, 2003.
- [10] Bradley Efron. Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis. *Journal of the American Statistical Association*, 99:465, 2004.
- [11] Bradley Efron. *Local False Discovery Rates*. Division of Biostatistics, Stanford University, 2005.
- [12] Bradley Efron. Doing Thousands of Hypothesis Tests at the Same Time. *Metron-International Journal of Statistics*, 65(1):3–21, 2007.
- [13] Bradley Efron. *Large-Scale Inference: Empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press, 2012.
- [14] Bradley Efron, Robert Tibshirani, John D Storey, and Virginia Tusher. Empirical Bayes Analysis of a Microarray experiment. *Journal of the American statistical association*, 96(456):1151–1160, 2001.

- [15] Felix Famoye and Karan P Singh. Zero inflated Generalized Poisson regression model with an application to domestic violence data. *Journal of Data Science*, 4(1):117–130, 2006.
- [16] Susanne Gschlößl and Claudia Czado. Modelling count data with overdispersion and spatial effects. *Statistical papers*, 49(3):531–552, 2008.
- [17] Pushpa Lata Gupta, Ramesh C Gupta, and Ram C Tripathi. Score test for Zero-inflated Generalized Poisson regression model. *Communications in Statistics-Theory and Methods*, 33(1):47–64, 2005.
- [18] Tatsuhiko Imaoka, Tomomi Okutani, Kazuhiro Daino, Daisuke Iizuka, Mayumi Nishimura, and Yoshiya Shimada. Overexpression of notch-regulated ankyrin repeat protein is associated with breast cancer cell proliferation. *Anticancer research*, 34(5):2165–2171, 2014.
- [19] A Jeanes, CJ Gottardi, and AS Yap. Cadherins and cancer: how does cadherin dysfunction promote tumor progression&quest. *Oncogene*, 27(55):6920–6929, 2008.
- [20] Jiashun Jin and T Tony Cai. Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons. *Journal of the American Statistical Association*, 102(478):495–506, 2007.
- [21] Harry Joe and Rong Zhu. Generalized Poisson distribution: the property of mixture of Poisson and comparison with negative binomial distribution. *Biometrical Journal*, 47(2):219–229, 2005.
- [22] Kim R Kampen. Membrane proteins: the key players of a cancer cell. *The Journal of membrane biology*, 242(2):69–74, 2011.
- [23] BM Golam Kibria. Applications of some discrete regression models for count data. *Pakistan Journal of Statistics and Operation Research*, 2(1), 2006.
- [24] Bernhard Klar. Bounds on Tail Probabilities of Discrete Distributions. *Probability in the Engineering and Informational Sciences*, 14:161–171, 4 2000.
- [25] Diane Lambert. Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14, 1992.
- [26] Rikke Leth-Larsen, Rikke Lund, Helle V Hansen, Anne-Vibeke Laenkholm, David Tarin, Ole N Jensen, and Henrik J Ditzel. Metastasis-related plasma membrane proteins of human breast cancer cells identified by comparative quantitative mass spectrometry. *Molecular & Cellular Proteomics*, 8(6):1436–1449, 2009.
- [27] Junan Li, Anjali Mahajan, and Ming-Daw Tsai. Ankyrin repeat: a unique motif mediating protein-protein interactions. *Biochemistry*, 45(51):15168–15178, 2006.
- [28] G. J. McLachlan and P. N. Jones. Fitting Mixture Models to Grouped and Truncated data via the EM Algorithm. *Biometrics*, 44:571–578, 1988.
- [29] Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, New York, NY, USA, 2005.
- [30] Rena Morita, Yoshihiko Hirohashi, Toshihiko Torigoe, Satoko INODA, Akari Takahashi, Tasuku Mariya, Hiroko Asanuma, Yasuaki Tamura, Tomohide Tsukahara, Takayuki Kanaseki, et al. Olfactory receptor family receptor, family 7, subfamily c, member 1 is a novel marker of colon cancer-initiating cells and is a potent target of immunotherapy. *Clinical Cancer Research*, pages clincanres–1709, 2016.

- [31] Nathan L Nehrt, Thomas A Peterson, DoHwan Park, and Maricel G Kann. Domain landscapes of somatic mutations in cancer. *BMC genomics*, 13(Suppl 4):S9, 2012.
- [32] Eva M Neuhaus, Weiyi Zhang, Lian Gelis, Ying Deng, Joachim Noldus, and Hanns Hatt. Activation of an olfactory receptor inhibits proliferation of prostate cancer cells. *Journal of Biological Chemistry*, 284(24):16218–16225, 2009.
- [33] Mariko Ohmori, Senji Shirasawa, Masanori Furuse, Koji Okumura, and Takehiko Sasazuki. Activated ki-ras enhances sensitivity of ceramide-induced apoptosis without c-jun nh2-terminal kinase/stress-activated protein kinase or extracellular signal-regulated kinase activation in human colon cancer cells. *Cancer research*, 57(21):4714–4717, 1997.
- [34] DoHwan Park, Junyong Park, Xiaosong Zhong, and Michel Sadelain. Estimation of empirical null using a mixture of normals and its use in local false discovery rate. *Computational Statistics & Data Analysis*, 55(7):2421–2432, 2011.
- [35] Giovanni Parmigiani, J Lin, Simina Boca, T Sjoblom, KW Kinzler, VE Velculescu, and B Vogelstein. Statistical methods for the analysis of cancer genome sequencing data. 2007.
- [36] Thomas A Peterson, Asa Adadey, Ivette Santana-Cruz, Yanan Sun, Andrew Winder, and Maricel G Kann. DMDM: Domain Mapping of Disease Mutations. *Bioinformatics*, 26(19):2458–2459, 2010.
- [37] Thomas A Peterson, Nathan L Nehrt, DoHwan Park, and Maricel G Kann. Incorporating Molecular and Functional context into the analysis and prioritization of human variants associated with cancer. *Journal of the American Medical Informatics Association*, 19(2):275–283, 2012.
- [38] Thomas A Peterson, DoHwan Park, and Maricel G Kann. A protein domain-centric approach for the comparative analysis of human and yeast phenotypically relevant mutations. *BMC genomics*, 14(3):1, 2013.
- [39] YN Phang and EF Loh. Zero inflated models for overdispersed count data. In *Proceedings of World Academy of Science, Engineering and Technology*, number 80, page 652. World Academy of Science, Engineering and Technology (WASET), 2013.
- [40] Katherine S Pollard, Merrill D Birkner, Mark J Van Der Laan, and Sandrine Dudoit. Test Statistics Null Distributions in Multiple Testing: Simulation studies and applications to Genomics. *Journal de la société française de statistique*, 146(1-2):77–115, 2005.
- [41] Eric K Rowinsky. Signal events: cell signal transduction and its inhibition in cancer. *The Oncologist*, 8(Supplement 3):5–17, 2003.
- [42] Guenhaël Sanz, Isabelle Leray, Aurélie Dewaele, Julien Sobilo, Stéphanie Lerondel, Stéphan Bouet, Denise Grébert, Régine Monnerie, Edith Pajot-Augy, and Lluís M Mir. Promotion of cancer cell invasiveness and metastasis emergence caused by olfactory receptor stimulation. *PloS one*, 9(1):e85110, 2014.
- [43] Richard Sever and Joan S Brugge. Signal transduction in cancer. *Cold Spring Harbor perspectives in medicine*, 5(4):a006098, 2015.
- [44] Sergey Sheetlin, Yonil Park, and John L Spouge. Objective method for estimating asymptotic parameters, with an application to sequence alignment. *Physical Review E*, 84(3):031914, 2011.
- [45] John D Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002.

- [46] Michael R Stratton. Exploring the Genomes of cancer cells: progress and promise. *science*, 331(6024):1553–1558, 2011.
- [47] Michael R Stratton, Peter J Campbell, and P Andrew Futreal. The Cancer Genome. *Nature*, 458(7239):719–724, 2009.
- [48] Masatoshi Takeichi. Cadherins in cancer: implications for invasion and metastasis. *Current opinion in cell biology*, 5(5):806–811, 1993.
- [49] Christos Tsatsanis and Demetrios A Spandidos. The role of oncogenic kinases in human cancer (review). *International journal of molecular medicine*, 5:583–590, 2000.
- [50] Shahid Ullah, Caroline F Finch, and Lesley Day. Statistical modelling for falls count data. *Accident Analysis & Prevention*, 42(2):384–392, 2010.
- [51] Esther Witsch, Michael Sela, and Yosef Yarden. Roles for growth factors in cancer progression. *Physiology*, 25(2):85–101, 2010.
- [52] Yinglin Xia, Dianne Morrison-Beedy, Jingming Ma, Changyong Feng, Wendi Cross, and Xin Tu. Modeling count outcomes from HIV risk reduction interventions: a comparison of competing statistical models for count responses. *AIDS research and treatment*, 2012, 2012.
- [53] Fan Yang, Evangelia Petsalaki, Thomas Rolland, David E Hill, Marc Vidal, and Frederick P Roth. Protein domain-level landscape of cancer-type-specific somatic mutations. *PLOS Comput Biol*, 11(3):e1004147, 2015.

Appendix:

E- Step:

At the $(p+1)th$ stage, the expectation $Q(\Theta; \Theta^{(p)})$ of the log-likelihood of the complete data specified in (9) can be computed conditional on the observed data \mathbf{y}_n and the current fit $\Theta^{(p)}$ for Θ .

$$\begin{aligned}
Q(\Theta; \Theta^{(p)}) &= n_0 \tau_{00}(\Theta^{(p)}) \log \eta + \sum_{j=0}^C n_j \tau_{1j}(\Theta^{(p)}) \log(1 - \eta) + \sum_{j=0}^C np_j(\Theta^{(p)}) \tau_{1j}(\Theta^{(p)}) \log(1 - \eta) \\
&\quad + (\log \lambda - \lambda) \sum_{j=0}^C n_j \tau_{1j}(\Theta^{(p)}) + (\log \lambda - \lambda) \sum_{j=C+1}^K np_j(\Theta^{(p)}) \tau_{1j}(\Theta^{(p)}) \\
&\quad + \sum_{j=0}^C n_j (j-1) \tau_{1j}(\Theta^{(p)}) \log(\lambda + \theta j) + \sum_{j=C+1}^K n(j-1) p_j(\Theta^{(p)}) \tau_{1j}(\Theta^{(p)}) \log(\lambda + \theta j) \\
&\quad - \left[\theta \sum_{j=0}^C j n_j \tau_{1j}(\Theta^{(p)}) + \theta \sum_{j=C+1}^K j np_j(\Theta^{(p)}) \tau_{1j}(\Theta^{(p)}) + constant \right]
\end{aligned}$$

M- Step:

In order to arrive at an estimate of $\Theta^{(p+1)}$ at the $(p+1)th$ stage, the goal is to maximize $Q(\Theta; \Theta^{(p)})$ with respect to Θ . The estimates of η, λ and θ obtained at the $(p+1)th$ stage are as follows:

$$\begin{aligned}
\eta^{(p+1)} &= \frac{n_0 \tau_{00}(\Theta^{(p)})}{n_0 \tau_{00}(\Theta^{(p)}) + \sum_{j=0}^C n_j \tau_{1j}(\Theta^{(p)}) + \sum_{j=C+1}^K np_j(\Theta^{(p)}) \tau_{1j}(\Theta^{(p)})} \\
\lambda^{(p+1)} &= \frac{\sum_{j=0}^C n_j [\tau_{1j}(\Theta^{(p)}) + (j-1) \tau_{2j}(\Theta^{(p)})] + \sum_{j=C+1}^K np_j(\Theta^{(p)}) [\tau_{1j}(\Theta^{(p)}) + (j-1) \tau_{2j}(\Theta^{(p)})]}{\sum_{j=0}^C n_j \tau_{1j}(\Theta^{(p)}) + \sum_{j=C+1}^K np_j(\Theta^{(p)}) \tau_{1j}(\Theta^{(p)})} \\
\theta^{(p+1)} &= \frac{\sum_{j=0}^C n_j (j-1) \tau_{3j}(\Theta^{(p)}) + \sum_{j=C+1}^K np_j (j-1) \tau_{3j}(\Theta^{(p)})}{\sum_{j=0}^C j n_j \tau_{1j}(\Theta^{(p)}) + \sum_{j=C+1}^K j np_j(\Theta^{(p)}) \tau_{1j}(\Theta^{(p)})}
\end{aligned}$$

where $\tau_{2j}(\Theta^{(p)}) = \frac{\lambda^{(p)}}{\lambda^{(p)} + \theta^{(p)} j}$ and $\tau_{3j}(\Theta^{(p)}) = \frac{\theta^{(p)} j}{\lambda^{(p)} + \theta^{(p)} j}$.

If the null distribution is modeled using Zero-Inflated Poisson distribution then the log likelihood $\ell(\eta, \lambda \mid \mathbf{x}_N)$ of the entire data vector is

$$n_0 \log \left(\eta + (1 - \eta) e^{-\lambda} \right) + \sum_{j=1}^C n_j \log(1 - \eta) \frac{\lambda^j e^{-\lambda}}{j!} + \sum_{j=C+1}^K n_j \log f(j; \cdot)$$

Following the same procedure, the E -Step at the $(p+1)th$ stage would yield

$$\begin{aligned}
Q(\Theta; \Theta^{(p)}) &= n_0 \tau_{00}(\Theta^{(p)}) \log \eta + \sum_{j=0}^C n_j \tau_{1j}(\Theta^{(p)}) \log(1 - \eta) + \sum_{j=0}^C np_j(\Theta^{(p)}) \tau_{1j}(\Theta^{(p)}) \log(1 - \eta) \\
&\quad + \log \lambda \sum_{j=0}^C j n_j \tau_{1j}(\Theta^{(p)}) + \log \lambda \sum_{j=C+1}^K j np_j(\Theta^{(p)}) \tau_{1j}(\Theta^{(p)}) \\
&\quad - \left[\lambda \sum_{j=0}^C n_j \tau_{1j}(\Theta^{(p)}) + \lambda \sum_{j=C+1}^K np_j(\Theta^{(p)}) \tau_{1j}(\Theta^{(p)}) + constant \right]
\end{aligned}$$

For the M -Step, the estimates of η and λ obtained at the $(p+1)th$ stage are as follows:

$$\begin{aligned}
\eta^{(p+1)} &= \frac{n_0 \tau_{00}(\Theta^{(p)})}{n_0 \tau_{00}(\Theta^{(p)}) + \sum_{j=0}^C n_j \tau_{1j}(\Theta^{(p)}) + \sum_{j=C+1}^K np_j(\Theta^{(p)}) \tau_{1j}(\Theta^{(p)})} \\
\lambda^{(p+1)} &= \frac{\sum_{j=0}^C j n_j \tau_{1j}(\Theta^{(p)}) + \sum_{j=C+1}^K j np_j(\Theta^{(p)}) \tau_{1j}(\Theta^{(p)})}{\sum_{j=0}^C n_j \tau_{1j}(\Theta^{(p)}) + \sum_{j=C+1}^K np_j(\Theta^{(p)}) \tau_{1j}(\Theta^{(p)})}
\end{aligned}$$

If the null distribution is modeled using Generalized Poisson distribution then the log likelihood $\ell(\lambda, \theta \mid \mathbf{x}_N)$ of the entire data is

$$\sum_{j=0}^C n_j \log \left(\frac{\lambda(\lambda + \theta j)^{j-1} e^{-\lambda - \theta j}}{j!} \right) + \sum_{j=C+1}^K n_j \log f(j; \cdot)$$

Unlike ZIGP, this model is not a mixture density so the procedure does not require the inclusion of latent variables. The E -Step would yield

$$\begin{aligned}
Q(\Theta; \Theta^{(p)}) &= (\log \lambda - \lambda) \sum_{j=0}^C n_j + (\log \lambda - \lambda) \sum_{j=C+1}^K np_j(\Theta^{(p)}) + \sum_{j=0}^C n_j (j-1) \log(\lambda + \theta j) \\
&\quad + \sum_{j=C+1}^K n(j-1) p_j(\Theta^{(p)}) \log(\lambda + \theta j) - \left[\theta \sum_{j=0}^C j n_j + \theta \sum_{j=C+1}^K j np_j(\Theta^{(p)}) + constant \right]
\end{aligned}$$

At the $(p+1)th$ stage, the M -Step would yield the estimates of λ and θ as follows:

$$\begin{aligned}
\lambda^{(p+1)} &= \frac{\sum_{j=0}^C n_j [1 + (j-1) \tau_{2j}(\Theta^{(p)})] + \sum_{j=C+1}^K np_j(\Theta^{(p)}) [1 + (j-1) \tau_{2j}(\Theta^{(p)})]}{\sum_{j=0}^C n_j + \sum_{j=C+1}^K np_j(\Theta^{(p)})} \\
\theta^{(p+1)} &= \frac{\sum_{j=0}^C n_j (j-1) \tau_{3j}(\Theta^{(p)}) + \sum_{j=C+1}^K n(j-1) p_j(\Theta^{(p)}) \tau_{3j}(\Theta^{(p)})}{\sum_{j=0}^C j n_j + \sum_{j=C+1}^K j np_j(\Theta^{(p)})}
\end{aligned}$$

Lastly, if f_0 is modeled using Poisson distribution then the log likelihood $\ell(\lambda \mid \mathbf{x}_N)$ of the entire data vector is

$$\sum_{j=0}^C n_j \log \left(\frac{\lambda^j e^{-\lambda}}{j!} \right) + \sum_{j=C+1}^K n_j \log f(j; \cdot)$$

Since this model is also not a mixture density then the procedure does not require the inclusion of zero-one indicator variables. The E -Step would yield

$$\begin{aligned} Q(\Theta; \Theta^{(p)}) &= \log \lambda \sum_{j=0}^C j n_j + \log \lambda \sum_{j=C+1}^K j n p_j(\Theta^{(p)}) \\ &\quad - \left[\lambda \sum_{j=0}^C n_j \tau_{1j}(\Theta^{(p)}) + \lambda \sum_{j=C+1}^K n p_j \tau_{1j}(\Theta^{(p)}) + constant \right] \end{aligned}$$

The M -Step would yield the estimate of λ at the $(p+1)th$ stage as follows:

$$\lambda^{(p+1)} = \frac{\sum_{j=0}^C j n_j + \sum_{j=C+1}^K j n p_j(\Theta^{(p)})}{\sum_{j=0}^C n_j + \sum_{j=C+1}^K n p_j(\Theta^{(p)})}$$

Supplementary Figures and Tables:

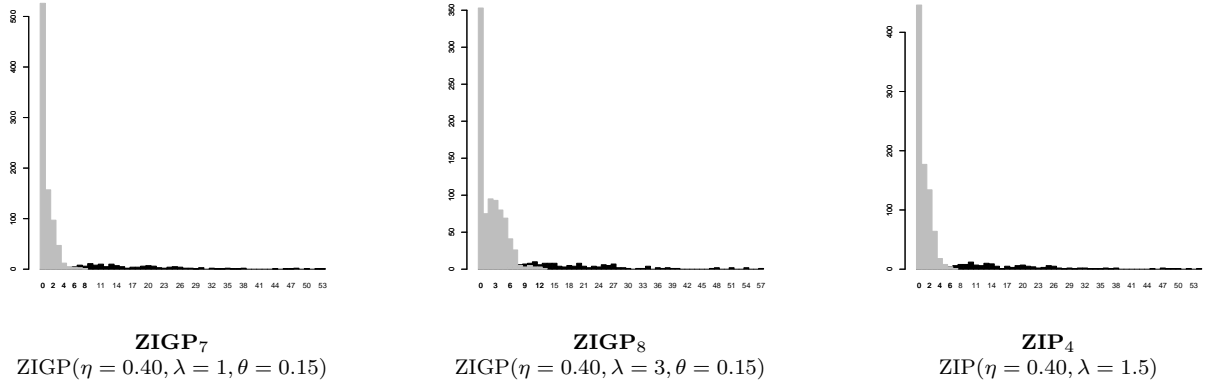


Figure 4. Histogram when the Non-null Distribution is Geometric($p = 0.08$) and $\pi_0 = 0.85$

Table 4. Numerical Comparison using C_1 as cut-off when the Non-null Distribution is Geometric($p = 0.08$), $\pi_0 = 0.85$ and $\alpha = 0.05$. The number in (\cdot) represents the standard deviation.

Null Distribution	Model for f_0	Two-Stage Procedure			One-Stage Procedure			Storey's FDR		
		R	\widehat{FDR}	\widehat{TPR}	R	\widehat{FDR}	\widehat{TPR}	R	\widehat{FDR}	\widehat{TPR}
ZIGP ₇	ZIGP	156.30 (12.02)	0.03821 (0.02781)	0.99995 (0.00091)	138.95 (12.70)	0.00700 (0.00784)	0.91765 (0.03306)	134.85 (11.25)	0.00440 (0.00596)	0.89345 (0.03003)
	ZIP	151.22 (12.36)	0.02186 (0.01488)	0.98405 (0.02298)	151.20 (12.37)	0.02185 (0.01489)	0.98399 (0.02307)	148.77 (12.26)	0.01788 (0.01366)	0.97222 (0.02825)
	GP	86.53 (71.46)	0.01001 (0.01464)	0.56913 (0.46522)	75.80 (63.43)	0.00381 (0.00672)	0.50808 (0.41601)	75.54 (61.58)	0.00199 (0.00471)	0.50418 (0.40734)
	P	190.45 (23.92)	0.20303 (0.08036)	1.00000 (0.00000)	190.45 (23.92)	0.20303 (0.08036)	1.00000 (0.00000)	175.45 (14.11)	0.14214 (0.04425)	1.00000 (0.00000)
ZIGP ₈	ZIGP	128.65 (21.91)	0.05204 (0.03398)	0.80943 (0.11523)	112.85 (20.03)	0.02111 (0.01892)	0.79575 (0.11091)	113.47 (18.41)	0.02096 (0.01722)	0.73885 (0.10381)
	ZIP	195.05 (12.87)	0.25656 (0.03248)	0.96497 (0.01509)	142.60 (21.31)	0.08532 (0.05996)	0.87498 (0.05326)	135.85 (15.46)	0.06452 (0.03620)	0.84347 (0.04738)
	GP	2.07 (2.49)	0.00000 (0.00000)	0.01419 (0.01797)	1.45 (2.78)	0.00000 (0.00000)	0.01419 (0.01797)	10.67 (5.91)	0.00000 (0.00000)	0.07234 (0.04276)
	P	496.36 (42.85)	0.69548 (0.03060)	1.00000 (0.00000)	496.36 (42.85)	0.69548 (0.03060)	1.00000 (0.00000)	391.38 (35.81)	0.61397 (0.03580)	1.00000 (0.00000)
ZIP ₄	ZIGP	152.37 (11.56)	0.01508 (0.01184)	0.99869 (0.00516)	135.25 (11.01)	0.00055 (0.00203)	0.89951 (0.02502)	135.46 (10.87)	0.00056 (0.00204)	0.90103 (0.02431)
	ZIP	144.35 (11.37)	0.00341 (0.00527)	0.95734 (0.01708)	142.47 (12.02)	0.00288 (0.00507)	0.94463 (0.02194)	141.56 (11.73)	0.00247 (0.00438)	0.93979 (0.03154)
	GP	146.60 (17.35)	0.00976 (0.01082)	0.96605 (0.08442)	128.21 (16.72)	0.00037 (0.00168)	0.85250 (0.08619)	126.81 (16.29)	0.00030 (0.00153)	0.84407 (0.08847)
	P	197.73 (30.67)	0.22736 (0.09507)	1.00000 (0.00000)	197.73 (30.67)	0.22736 (0.09507)	1.00000 (0.00000)	163.45 (14.77)	0.07836 (0.04829)	1.00000 (0.00000)

Table 5. Numerical Comparison using C_2 as cut-off, when the Non-null Distribution is Geometric($p = 0.08$), $\pi_0 = 0.85$ and $\alpha = 0.05$. The number in (\cdot) represents the standard deviation.

Null Distribution	Model for f_0	Two-Stage Procedure			One-Stage Procedure			Storey's FDR		
		R	\widehat{FDR}	\widehat{TPR}	R	\widehat{FDR}	\widehat{TPR}	R	\widehat{FDR}	\widehat{TPR}
ZIGP ₇	ZIGP	158.52 (11.85)	0.05181 (0.02711)	1.00000 (0.00000)	134.86 (11.49)	0.00455 (0.00627)	0.89319 (0.02918)	134.00 (10.86)	0.00395 (0.00550)	0.88832 (0.02645)
	ZIP	151.00 (12.33)	0.02163 (0.01490)	0.98292 (0.02383)	150.94 (12.38)	0.02158 (0.01494)	0.98264 (0.02437)	148.78 (12.22)	0.01803 (0.01396)	0.97213 (0.02868)
	GP	86.48 (71.43)	0.01003 (0.01467)	0.56873 (0.46501)	75.74 (63.40)	0.00378 (0.00670)	0.50698 (0.41646)	75.15 (61.91)	0.00220 (0.00489)	0.50151 (0.40962)
	P	190.45 (23.92)	0.20303 (0.08036)	1.00000 (0.00000)	190.45 (23.92)	0.20303 (0.08036)	1.00000 (0.00000)	175.45 (14.11)	0.14214 (0.04425)	1.00000 (0.00000)
ZIGP ₈	ZIGP	91.67 (56.87)	0.03038 (0.03400)	0.58328 (0.35421)	80.33 (49.98)	0.01183 (0.01687)	0.57384 (0.34786)	83.39 (46.34)	0.01184 (0.01584)	0.54606 (0.29789)
	ZIP	195.05 (12.87)	0.25656 (0.03248)	0.96497 (0.01509)	136.32 (27.85)	0.07661 (0.06723)	0.83890 (0.10124)	130.62 (20.64)	0.05678 (0.04213)	0.81515 (0.08111)
	GP	1.05 (0.23)	0.00000 (0.00000)	0.00703 (0.00162)	0.00 (0.00)	0.00000 (0.00000)	0.00703 (0.00162)	4.48 (1.58)	0.00000 (0.00000)	0.02989 (0.01052)
	P	496.36 (42.85)	0.69548 (0.03060)	1.00000 (0.00000)	496.36 (42.85)	0.69548 (0.03060)	1.00000 (0.00000)	391.38 (35.81)	0.61397 (0.03580)	1.00000 (0.00000)
ZIP ₄	ZIGP	152.73 (11.44)	0.01742 (0.01475)	0.99870 (0.00789)	131.27 (12.02)	0.00036 (0.00159)	0.87284 (0.03779)	129.89 (11.40)	0.00032 (0.00152)	0.86435 (0.04224)
	ZIP	144.35 (11.37)	0.00341 (0.00527)	0.95734 (0.01708)	140.70 (12.46)	0.00234 (0.00471)	0.93476 (0.03069)	139.93 (12.11)	0.00197 (0.00391)	0.92932 (0.03658)
	GP	147.50 (17.46)	0.01106 (0.01080)	0.97085 (0.08649)	127.70 (16.84)	0.00036 (0.00166)	0.84953 (0.08692)	126.42 (16.41)	0.00029 (0.00151)	0.84149 (0.08957)
	P	197.73 (30.67)	0.22736 (0.09507)	1.00000 (0.00000)	197.73 (30.67)	0.22736 (0.09507)	1.00000 (0.00000)	163.45 (14.77)	0.07836 (0.04829)	1.00000 (0.00000)

Figure 5. Histogram when the Non-null Distribution is Binomial($n = 250, p = 0.20$) and $\pi_0 = 0.70$

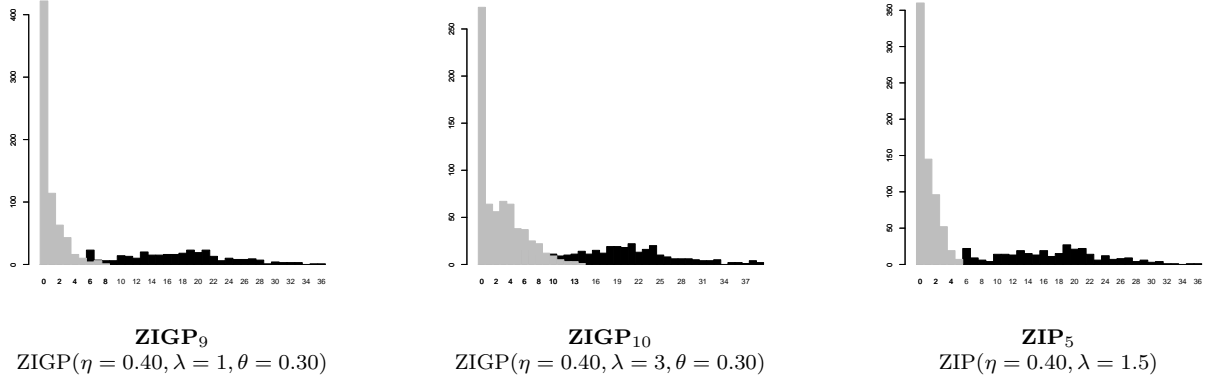


Table 6. Numerical Comparison using C_1 as cut-off when the Non-null Distribution is Binomial($n = 250, p = 0.20$), $\pi_0 = 0.70$ and $\alpha = 0.05$. The number in (\cdot) represents the standard deviation.

Null Distribution	Model for f_0	Two-Stage Procedure			One-Stage Procedure			Storey's FDR		
		R	\widehat{FDR}	\widehat{TPR}	R	\widehat{FDR}	\widehat{TPR}	R	\widehat{FDR}	\widehat{TPR}
ZIGP ₉	ZIGP	316.41 (15.24)	0.05152 (0.02046)	0.99987 (0.00248)	283.86 (14.30)	0.01551 (0.00882)	0.93132 (0.01549)	288.81 (14.41)	0.02115 (0.00994)	0.94208 (0.01644)
	ZIP	311.60 (15.96)	0.04190 (0.01461)	0.99472 (0.01495)	311.44 (16.15)	0.04171 (0.01489)	0.99472 (0.01495)	307.52 (17.17)	0.03854 (0.01521)	0.98507 (0.02425)
	GP	1.22 (0.55)	0.00000 (0.00000)	0.00408 (0.00184)	0.00 (0.00)	0.00000 (0.00000)	0.00180 (0.00221)	4.01 (2.40)	0.00000 (0.00000)	0.01335 (0.00796)
	P	377.83 (20.84)	0.20463 (0.03695)	1.00000 (0.00000)	377.83 (20.84)	0.20463 (0.03695)	1.00000 (0.00000)	345.32 (14.79)	0.13101 (0.01855)	1.00000 (0.00000)
ZIGP ₁₀	ZIGP	237.25 (90.17)	0.04893 (0.03985)	0.74353 (0.27104)	230.78 (87.49)	0.04185 (0.03263)	0.74340 (0.27096)	248.22 (67.64)	0.04711 (0.03405)	0.78263 (0.19811)
	ZIP	375.04 (19.83)	0.22727 (0.02554)	0.96514 (0.01382)	328.08 (39.07)	0.13961 (0.06537)	0.94225 (0.02774)	320.20 (27.39)	0.12648 (0.04568)	0.92887 (0.02494)
	GP	1.20 (0.50)	0.00000 (0.00000)	0.00399 (0.00165)	0.00 (0.00)	0.00000 (0.00000)	0.00399 (0.00165)	10.30 (5.04)	0.00000 (0.00000)	0.03430 (0.01658)
	P	611.10 (34.36)	0.50773 (0.03149)	1.00000 (0.00000)	611.10 (34.36)	0.50773 (0.03149)	1.00000 (0.00000)	577.55 (31.39)	0.47925 (0.03244)	1.00000 (0.00000)
ZIP ₅	ZIGP	301.86 (14.07)	0.00596 (0.00489)	1.00000 (0.00000)	279.17 (13.54)	0.00021 (0.00090)	0.93021 (0.01482)	283.37 (14.02)	0.00082 (0.00179)	0.94361 (0.01624)
	ZIP	287.81 (14.89)	0.00186 (0.00300)	0.95733 (0.01938)	285.50 (15.47)	0.00150 (0.00293)	0.95699 (0.01939)	287.33 (14.75)	0.00171 (0.00275)	0.95592 (0.02013)
	GP	258.70 (88.48)	0.00370 (0.00471)	0.86037 (0.29058)	242.07 (81.70)	0.00015 (0.00074)	0.80723 (0.26859)	248.86 (79.44)	0.00042 (0.00125)	0.83079 (0.26233)
	P	338.33 (26.65)	0.10956 (0.05600)	1.00000 (0.00000)	338.33 (26.65)	0.10956 (0.05600)	1.00000 (0.00000)	322.21 (16.92)	0.06800 (0.02948)	1.00000 (0.00000)

Table 7. Numerical Comparison using C_2 as cut-off when the Non-null Distribution is Binomial($n = 250, p = 0.20$), $\pi_0 = 0.80$ and $\alpha = 0.05$. The number in (\cdot) represents the standard deviation.

Null Distribution	Model for f_0	Two-Stage Procedure			One-Stage Procedure			Storey's FDR		
		R	\widehat{FDR}	\widehat{TPR}	R	\widehat{FDR}	\widehat{TPR}	R	\widehat{FDR}	\widehat{TPR}
ZIGP ₉	ZIGP	314.70 (15.04)	0.04721 (0.01780)	0.99907 (0.00898)	282.66 (14.18)	0.01445 (0.00794)	0.92840 (0.01745)	286.89 (14.61)	0.01901 (0.00978)	0.93786 (0.01796)
	ZIP	312.34 (15.43)	0.04253 (0.01447)	0.99648 (0.01225)	312.23 (15.58)	0.04239 (0.01468)	0.99648 (0.01225)	309.36 (16.99)	0.03991 (0.01486)	0.98951 (0.02147)
	GP	1.22 (0.55)	0.00000 (0.00000)	0.00408 (0.00184)	0.00 (0.00)	0.00000 (0.00000)	0.00041 (0.00123)	2.23 (1.52)	0.00000 (0.00000)	0.00741 (0.00504)
	P	377.83 (20.84)	0.20463 (0.03695)	1.00000 (0.00000)	377.83 (20.84)	0.20463 (0.03695)	1.00000 (0.00000)	345.32 (14.79)	0.13101 (0.01855)	1.00000 (0.00000)
ZIGP ₁₀	ZIGP	97.34 (130.74)	0.02279 (0.03934)	0.30289 (0.40247)	92.95 (127.11)	0.02126 (0.03328)	0.30282 (0.40237)	139.75 (100.95)	0.02109 (0.03428)	0.44630 (0.30743)
	ZIP	375.16 (20.02)	0.22740 (0.02568)	0.96525 (0.01387)	334.29 (40.71)	0.15174 (0.06618)	0.94528 (0.03121)	324.19 (28.79)	0.13421 (0.04682)	0.93174 (0.02613)
	GP	1.20 (0.50)	0.00000 (0.00000)	0.00399 (0.00165)	0.00 (0.00)	0.00000 (0.00000)	0.00399 (0.00165)	11.26 (4.89)	0.00000 (0.00000)	0.03751 (0.01598)
	P	611.10 (34.36)	0.50773 (0.03149)	1.00000 (0.00000)	611.10 (34.36)	0.50773 (0.03149)	1.00000 (0.00000)	577.55 (31.39)	0.47925 (0.03244)	1.00000 (0.00000)
ZIP ₅	ZIGP	301.12 (14.39)	0.00611 (0.00556)	0.99739 (0.01144)	275.43 (14.21)	0.00011 (0.00064)	0.91829 (0.01818)	279.84 (13.75)	0.00028 (0.00102)	0.93237 (0.01573)
	ZIP	288.73 (15.34)	0.00219 (0.00343)	0.96005 (0.02126)	285.97 (16.18)	0.00176 (0.00340)	0.95792 (0.02223)	287.75 (15.28)	0.00193 (0.00314)	0.95708 (0.02300)
	GP	221.18 (128.96)	0.00372 (0.00472)	0.73562 (0.42671)	204.70 (119.76)	0.00018 (0.00081)	0.68316 (0.39767)	209.69 (119.83)	0.00042 (0.00125)	0.70059 (0.39860)
	P	338.33 (26.65)	0.10956 (0.05600)	1.00000 (0.00000)	338.33 (26.65)	0.10956 (0.05600)	1.00000 (0.00000)	322.21 (16.92)	0.06800 (0.02948)	1.00000 (0.00000)